Psychometrics, Test Theory, and the Latent Factors Model

Petr Blahuš • Bruce L. Brown
C. Victor Bunderson • Joseph A. Olsen

WILEY

Psychometrics, Test Theory, and the Latent Factors Model

Psychometrics, Test Theory, and the Latent Factors Model

Petr Blahuš

Bruce L. Brown
Brigham Young University
Provo, United States

C. Victor Bunderson
Brigham Young University
Provo, United States

Joseph A. Olsen
Brigham Young University
Provo, United States

WILEY

2026 John Wiley & Sons, Inc. Published 2026 by John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

The manufacturer's authorized representative according to the EU General Product Safety Regulation is Wiley-VCH GmbH, Boschstr. 12, 69469 Weinheim, Germany, e-mail: Product_Safety@wiley.com.

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and the authors have used their best efforts in preparing this work, including a review of the content of the work, neither the publisher nor the authors make any representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data Applied for:

Hardback ISBN: 9781119312178

Cover Design: Wiley
Cover Image: © Michael H/Getty Images

Set in 9.5/12.5pt STIXTwoText by Straive, Pondicherry, India

Contents

About the Authors *xvii*
Foreword *xviii*
Preface *xx*
Acknowledgements *xxvii*
About the Companion Website *xxxii*

1 The Psychometric Quest *1*
1.1 Quantification in Psychometrics: Historical Origins *3*
1.1.1 (§1) Why Are Mathematical Methods Central to the Empirical Sciences? Why Are Statistical Methods Used in the Behavioral Sciences? *3*
1.1.2 (§2) The Specific Mission of Statistical Methods in the Behavioral Sciences *4*
1.1.3 (§3) Why Do We Use *Behavioral Statistics* and Not Just a Branch of Applied Mathematics? *4*
1.1.4 (§4) Some Remarks on the Historical Relationships Between Psychology and Statistics: The Development of the Correlation Coefficient and Least Squares Regression *5*
1.1.4.1 The Correlation Coefficient and Least Squares Regression *6*
1.1.4.2 Correlation, Regression, and Heritability *7*
1.2 Variables, Scales, and Data *9*
1.2.1 (§5) *On Terminology I*: Attributes with Categories, Variables with Values, Tests with Scores, and What About "Data"? *9*
1.2.1.1 Variables and Data *9*
1.2.1.2 Scientific Rigor, Representation, and Stochastic Modeling *11*
1.2.2 (§6) Measurement Scales and the Representational Theory of Measurement (RTM) *12*
1.2.2.1 The Ferguson Committee *12*
1.2.2.2 Stevens' Representational Theory of Measurement (RTM) and of Scale Types *13*
1.2.2.3 Stevens' Four Measurement Scales *13*
1.2.2.4 Foundations of Measurement, the "Received View" of RTM *16*
1.2.2.5 The Contrast Between Foundations of Measurement and the Science of Applied Psychometrics *19*
1.2.2.6 Metrology and Measurement Units *21*
1.2.2.7 Measurability in the Behavioral and Social Sciences *23*
1.2.3 (§7) *Scientific* and *Stochastic* Variables: When May "Data" Be Treated By Statistical and Test Theory Methods? *24*
1.3 Statistical Methods: Location, Dispersion, and Statistical Inference *25*
1.3.1 (§8) Group Level of an Attribute—The Mean Location of the Group's Scores on an Interval Scale *25*

1.3.2 (§9) Comparing Median vs. Mean: Which One Is "Fair"? *28*
1.3.3 (§10) A Psychometric Look at the Mean $\overline{X}$, Item "Difficulty," and p *28*
1.3.4 (§11) Intra-group and Inter-individual Variability—*Dispersion* of Scores *31*
1.3.5 (§12) A Small Numerical Example, Finally *33*
1.3.6 (§13) Does the Larger Head Circumference in the Sample of Sons Than Fathers Hold for the Whole Population? A Preliminary Sidestep to Statistical Inference *35*
1.3.6.1 An Estimate of the Probable Error of a Point Estimate *38*
1.3.6.2 Classical Probability Theory and the Binomial Distribution: Creating a Simple Stochastic Model *40*
1.3.6.3 The Binomial Distribution as a Simple and Clear Example of a Sampling Distribution of Proportions *46*
1.3.6.4 Elaboration of Higher Level Calculations for the Binomial Distribution *49*
1.3.6.5 Multinomial Demonstrations of Two Fundamental Principles of Statistical Inference *52*
1.4 Summary *56*
1.5 Study Questions *56*
References *62*

2 Scores and Their Distributions *65*
2.1 Quantification in Distributions: The Standard Deviation and Standard Scores *65*
2.1.1 (§14) A Psychometric Look at the Variance and SD: "Discriminability" of an Item *66*
2.1.1.1 Inappropriate Scoring *67*
2.1.1.2 The Standard Deviation and Binary Tests *67*
2.1.2 (§15) Using SD for Comparing the Incomparable: Percents vs. Z-Scores *68*
2.1.2.1 Norms and Baselines *68*
2.1.2.2 Using Percents with Norms and Baselines *69*
2.1.2.3 Comparing Scores Through Standardizing *70*
2.1.3 (§16) Density of Scores Along the Scale: The Normal Distribution of Frequency or Probability, the N-Law *72*
2.1.3.1 The Inverted Ogive *72*
2.1.3.2 The "Bell Curve" *73*
2.1.4 (§17) The Bell Curve or Normal Curve Distribution Function $p(Z)$ *73*
2.1.4.1 The Rationale Underlying the Simple Approximate Probability Distribution Equation *77*
2.1.4.2 The Need for a Cumulative Density Function *77*
2.1.5 (§18) The Ogive Curve or Normal Curve Cumulative Distribution Function $N(Z)$ *78*
2.1.6 (§19) Percentile Position of a Score and the N-Law of Distribution *80*
2.1.7 (§20) Are Z-Scores and Percentiles Fair? *81*
2.2 Combining and Using Standard Scores *82*
2.2.1 (§21) More Convenient Scores Derived from Z: The Deviation IQ, T-Score, and Stanine *83*
2.2.2 (§22) What If the Scores Do Not Follow the N-Law? A Skewed Bell Curve *84*
2.2.2.1 Floor and Ceiling Effects: J-Shaped and L-Shaped Distributions *86*
2.2.2.2 Variables with Inherently Non-normal Distributions *87*
2.2.2.3 Population Selection and Extreme Groups *88*
2.2.3 (§23) Composite Z-Scores of Different Tests: Adding Apples and Oranges? *88*
2.2.4 (§24) Mutual Compensation, and Differential Importance of Sub-tests in the Composite *90*
2.3 Summary *91*
2.4 Study Questions *93*
References *96*

3 Validity, Explanation, and Constructs *97*
3.1 Validity and Validation *99*
3.1.1 (§25) The Validity Dialectic *99*
3.1.1.1 Professional Standards *102*
3.1.1.2 The Validity Concept and the *Validation Argument* Today *103*
3.1.1.3 Measurement Instruments are an Inseparable Part of the Metrology Enterprise *107*
3.1.1.4 Measurement in Any Field of Science or Technology *108*
3.1.1.5 Summary: Validity Theory is Unsettled Historically and Theoretically and Needs Unification Again *110*
3.2 Terms and Tools in External Validity *111*
3.2.1 (§26) What is Validity? External, Internal, Concurrent, Predictive, and Practical Utility *112*
3.2.1.1 Standardized Composite Scores *112*
3.2.1.2 Criterion-Related Validity *113*
3.2.2 (§27) Is the Error of Estimating the Criterion Sufficiently Small for Some Test Uses? *114*
3.2.2.1 Composite Validity *115*
3.2.2.2 Variability and Validity *115*
3.2.3 (§28) Differential Weighting of Subtests: Composite Validity as Multiple Correlation *115*
3.2.3.1 Weighted Composite Validity *116*
3.2.3.2 Maximal Validity Via Multiple Correlation *116*
3.2.3.3 Multiple Correlation to Shorten a Test *117*
3.2.4 (§29) A Criterion Score is Best Estimated by a Particular Composite, Multiple Regression *118*
3.2.4.1 Comparing Several Methods for Constructing Composite Scores *118*
3.2.4.2 Multiple Regression with Raw Scores *119*
3.2.4.3 Standardized Multiple Regression *119*
3.2.4.4 Regression Calibration *120*
3.2.4.5 Error in Regression Prediction *120*
3.2.4.6 Remarks on Interpreting the Magnitude of Multiple *R* *120*
3.3 Explanation Requires More Than Correlation *121*
3.3.1 (§30) *Scientific Explanation and Prediction*: Multiple Correlation and Regression as Its Statistical Metaphor (i) *122*
3.3.1.1 "Explained" Variance in Regression *122*
3.3.1.2 Assessing Imprecision in Regression *123*
3.3.1.3 Fisher's Information and Efficiency Measures *125*
3.3.2 (§31) Correlation vs. Causation: Their Respective Roles in Scientific Explanation *125*
3.3.2.1 Experimental Manipulation and Passive Observation *126*
3.3.2.2 The Counterfactual Condition and a Thought Experiment *127*
3.3.3 (§32) The "Cleaned Off" Incremental Validity of a Subtest: Partial Correlation *127*
3.3.3.1 The Dilemma of Internal and External Validity *127*
3.3.3.2 Partial and Spurious Correlation *128*
3.3.3.3 Calculating Partial Correlation *129*
3.3.4 (§33) Scientific *Explanation of a Relationship* Between Attributes—Between Whole *Variables*: Partial Uncorrelatedness as Its *Statistical Metaphor (ii)* *129*
3.3.4.1 Partialing Via Sampling Stratification *130*
3.3.4.2 Partialing Via Statistical Control *130*
3.3.4.3 Local Uncorrelatedness and Falsifiability *130*
3.4 Internal Construct Validity and Latent Factors *132*
3.4.1 (§34) If the Explanatory Variable is Unobservable, *Construct* a Latent Factor (Construct Validity) *132*

3.4.1.1 Latent Variables and Local Uncorrelatedness *132*
3.4.1.2 A Latent Common Factor Model *134*
3.4.2 (§35) What are the Behavioral "Constructs" Behind the Scenes? The Theoretical Concepts of Scientific Theory *134*
3.4.2.1 Theoretical Concept Formation and Latent Variable Modeling *135*
3.4.2.2 Latent Variable Modeling and the Principle of Partial-Local Uncorrelatedness *135*
3.4.3 (§36) Causal Effects vs. Independent/Dependent Variables: A Systems Theory View *136*
3.4.3.1 Causal Variables and Effect Variables *137*
3.4.3.2 Descriptive Associations and Causal Relationships *138*
3.4.3.3 Systems Flowchart vs. Path Diagram *138*
3.4.4 (§37) Theoretical Concepts as Active Subsystems Hidden in the S ⇒ R Black Box: Causal and Propensity Interpretation of Ability *139*
3.4.5 (§38) Theoretical Concepts as *Attributes* Under *Indirect Observation*, with Their Indirect Measurement Through the Manifest-latent Link *142*
3.4.5.1 Manifest and Latent Variables *142*
3.4.5.2 Reading Comprehension Ability as a Formalizable Theoretical Construct *144*
3.4.6 (§39) On Terminology 2: Mishmash Around Constructs *146*
3.4.6.1 Formalized (Mathematical) and Unformalized (Scientific) Variables *147*
3.5 Summary *147*
3.6 Study Questions *148*
References *150*

4 Measurement Error and Classical Test Theory *153*
4.1 Introduction and Overview of "Unfinished Business" from Chapter 3 *154*
4.1.1 (§40) Domain Theory and Evidence-Centered Design *154*
4.1.1.1 Measurement Instruments and Their Design and Development *154*
4.1.1.2 Domain Theory *155*
4.1.1.3 Assessment in Relation to Measurement *157*
4.1.1.4 Resources for Design and Development *158*
4.1.1.5 How to Access Documentation and Resources for ECD and PADI *160*
4.2 Measurement Functions that Provide Associative Links Between the Empirical World and the Informational World *160*
4.2.1 (§41) The Gauss Model for Replications *164*
4.2.1.1 Replications and Random Errors *164*
4.2.2 (§42) On Direct, Derived, and Indirect Measurement Across the Sciences *164*
4.2.2.1 What Is Direct Measurement and How Does It Make Indirect Measurement Possible? *165*
4.2.2.2 Concatenation vs Instrumentation's Indirect Results *165*
4.2.2.3 Derived Measurement *166*
4.2.3 (§43) The *Specific True Score* τ of a Separate Test/Indicator *167*
4.2.3.1 Test–Retest Method *168*
4.2.3.2 Observed Score, True Score, and Error *168*
4.2.4 (§44) The Test as a Measuring Device Compared to the *Precision* of Measurement in Physics and Technology *169*
4.2.4.1 Imprecision and Relative Error *169*
4.2.4.2 European Governmental Standards and Guidelines for Defining Precision *171*
4.3 Reliability and Classical Test Theory *171*
4.3.1 (§45) Reliability vs. Precision *172*
4.3.2 (§46) Reliability of a Test, Efficiency of Measurement, and the Information Gained *173*

4.3.2.1 Fisher's Information and Efficiency *174*
4.3.2.2 Shannon's Entropy *174*
4.3.3 (§47) How to Recover the Unknown True τ? Randomness as the Fundamental Principle in Classical Test Theory *175*
4.3.3.1 Parallel Forms *176*
4.3.3.2 An Application of CT Theory *176*
4.3.3.3 The Expectation Principle *177*
4.3.3.4 The Randomness Principle *178*
4.3.4 (§48) Reliability of τ in Terms of the Platonic Conception of CTT *179*
4.3.4.1 Four Axioms of Classical Test Theory *179*
4.3.4.2 Using Classical Test Theory *180*
4.3.4.3 Some Practical Applications of Classical Test Theory *181*
4.3.4.4 Problems in Applying CT Theory to the Reliability of Derived Units *182*
4.4 Summary *182*
4.5 Study Questions *185*
References *189*

5 Content Validity, Item Response Theory, and Latent Variables *191*
5.1 Design and Initial Validation of Latent Factors and Domain Models *192*
5.1.1 (§49) A Dissertation-based Example of a Model of Progressive Attainment *192*
5.1.1.1 Focus on Internal Aspects of Validity *193*
5.1.1.2 Multiple Revisions Required *193*
5.1.1.3 Isolating the "FORE" Latent Factor from the "Word Knowledge" Latent Factor *193*
5.1.1.4 Other Models of Progressive Attainment *194*
5.2 Content Validity and Item Response Theory *195*
5.2.1 (§50) Causation and Content Validity *195*
5.2.1.1 Validating the Content Aspect of Validity *195*
5.2.1.2 Causal Explanations *196*
5.2.1.3 More on Content Aspects of Validity *197*
5.2.2 (§51) Content Validity and Content Homogeneity of Indicators/Subtests *198*
5.2.2.1 Content and Construct Validity *198*
5.2.2.2 Indexes of Expert Agreement *200*
5.2.2.3 Homogeneity *201*
5.2.2.4 Content Homogeneity vs. Unidimensionality *201*
5.2.2.5 Linear Factor Analysis of Categorical Items and "Difficulty Factors" *202*
5.2.3 (§52) Some Aspects of Item Response Theory and Computer-Adaptive Testing *204*
5.2.3.1 Item Characteristic Curves *204*
5.2.3.2 The Wright Map *204*
5.2.3.3 Computer-adaptive Testing: The High-jumping Metaphor *206*
5.2.3.4 Implementation Examples from the High-jumping Metaphor *207*
5.3 Latent Variable Modeling *207*
5.3.1 (§53) Generic Score of Homogeneous Indicators: Rough Approximations Without Latent Variables *208*
5.3.1.1 Test Blueprints—The Opposite of Homogeneity *209*
5.3.1.2 Formula Score *210*
5.3.1.3 Item Analysis and Factor Analysis *211*
5.3.2 (§54) The Types of Manifest and Latent Variables in Latent Variable Modeling: Values, Scales, and Probability Distributions *211*
5.3.2.1 Objects to be Modeled *211*

5.3.2.2 Level of Quantification *212*
5.3.2.3 Latent Factors, Latent Traits, and Latent Abilities *214*
5.3.2.4 Latent Class Models *215*
5.3.2.5 General Latent Variable Modeling *216*
5.3.2.6 Uncorrelatedness and Probabilistic Local Independence *217*
5.4 Summary *218*
5.4.1 Pitfalls to Avoid in So-called *Content Validation* *219*
5.4.2 Latent Variables and Factors *219*
5.5 Study Questions *220*
References *222*

6 The Latent Common Factor Model *225*
6.1 Latent Variables and the Common Factor Model *226*
6.1.1 (§55) Representation Conditions of Measurement and Correspondence Rules for Latent Variable Modeling *226*
6.1.1.1 Semantic Correspondence Rules *226*
6.1.1.2 Syntactic Correspondence Rules *227*
6.1.2 (§56) The Latent Common Factor Prototype of Latent Variable Modeling *230*
6.1.3 (§57) How to Calculate Loadings and Residual Correlations in the Factoring Process *232*
6.1.3.1 Applying the Latent Common Factor Model *232*
6.1.3.2 Centroid Method of Factoring *234*
6.1.3.3 Comparison of Three Models: The "Merely Historical Centroid Model," the PC Model, and the ML Model *235*
6.1.3.4 Fundamental Factor Theorem *238*
6.1.4 (§58) Maximize the Fit of the Model and Minimize the Discrepancy Index to Achieve Local Independence *239*
6.1.4.1 Least Squares Fitting and Model Fit Assessment *239*
6.1.4.2 Relative Model Fit *240*
6.1.4.3 Unidimensionality and Content Homogeneity *241*
6.1.4.4 Weighted Least Squares (WLS) Fitting *241*
6.1.4.5 Maximum Likelihood (ML) Fitting *242*
6.2 Factor Scores *243*
6.2.1 (§59) Estimating Each Subject's Latent Score *f* to Measure a Trait or Characteristic *243*
6.2.1.1 Estimating Factor Scores *244*
6.2.1.2 Factor Scores and Latent Values *246*
6.2.1.3 Generic True Scores and the Latent Common Factor Model *247*
6.2.2 (§60) Communality, the Variance Shared Among Indicators *247*
6.2.2.1 Communality, Specificity, and Unreliability *248*
6.2.2.2 Communality and the Squared Multiple Correlation (SMC) with the Other Indicators *249*
6.2.3 (§61) An Example Regression Estimate of *f* and the Decomposition of Indicator Variance *250*
6.2.3.1 The Tapping Test *251*
6.2.3.2 Test–Retest Analysis *251*
6.2.4 (§62) Generic Reliability and the "Model–Remodel" Assessment *253*
6.3 Summary *256*
6.4 Study Questions *257*
References *274*

7 Analysis of Multiple Factors *277*
7.1 Modeling Multiple and Hierarchical Constructs *277*
7.1.1 (§63) Multiconstruct Theories Including Two-Layer Structures *278*
7.1.1.1 Bi-factor Models *278*
7.1.1.2 Hierarchical Models *278*
7.1.1.3 Dimensionality and Higher-Order Constructs *280*
7.1.1.4 Constructs with Multiple Layers *280*
7.1.2 (§64) Homogeneity and Dimensionality in One-Layer Theories of Abilities: *Independent Cluster Structure (ICS)* Compared to *Simple Structure* *282*
7.1.2.1 Independent Cluster Structure *282*
7.1.2.2 Factor Rotation and Simple Structure *284*
7.1.3 (§65) Unidimensionality vs. Multidimensionality on One Layer, an Example *285*
7.2 Confirmatory and Exploratory *287*
7.2.1 (§66) The Confirmatory Approach *287*
7.2.2 (§67) Confirmatory vs. Exploratory Simple Structure *288*
7.2.2.1 Simple Structure *290*
7.2.2.2 Exploratory Common Factor Analysis *291*
7.2.2.3 Factor Rotation *292*
7.2.3 (§68) Stronger vs. Weaker Support for Content Homogeneity: By Unidimensionality or ICS *294*
7.2.3.1 Higher-Order Factor Analysis *295*
7.2.3.2 Equivocal Indicators and Factorial Complexity *296*
7.2.4 (§69) "Hidden Equivocality" in ICS Structures and the So-Called Frame Problem *296*
7.2.4.1 Indicator Multidimensionality *298*
7.2.4.2 The "Frame Problem" *299*
7.2.5 (§70) Dimensionality by the LC Factor *Model* vs. PCA "Factor *Analysis*"—Example *300*
7.2.5.1 Explaining Correlations vs. Explaining Variance-1 *300*
7.2.5.2 Meaninglessness of "Communalities" in PCA-2 *303*
7.2.5.3 What "Variance" Is Explained by PCA-3 *303*
7.2.5.4 Deterministic Component Scores vs. Factor Score Indeterminacy *305*
7.3 Summary *306*
7.4 Study Questions *307*
References *313*

8 Metrology: Scientific, Applied, and Legal: An Introduction with Implications for Human and Social Sciences *315*
8.1 Purposes of Metrology, the Science of Measurement *316*
8.2 Brief History of International Metrology and Its Relevance to Human Sciences *318*
8.2.1 The Ferguson Committee *320*
8.2.2 Why Is Metrology Missing in the Measurement Literature of the Human Sciences? *322*
8.3 Scientific Metrology *322*
8.3.1 International Scientific Metrology *323*
8.3.2 Why Was International Metrology Day, 20 May 2019, so Significant? *325*
8.3.3 Why Replace Physical Artifacts with Fundamental Constants of Nature? *325*
8.3.4 Examples: How the First Three SI Base Units Are Defined Scientifically *325*
8.3.4.1 The Second *326*
8.3.4.2 The Metre *326*
8.3.4.3 The Kilogram *326*

8.4 Applied and Legal Metrology *327*
8.4.1 Applied Metrology Purposes and Key Organizations *328*
8.4.2 Legal Metrology Purposes and Key Organizations *328*
8.5 A Manifesto for Establishing International Metrology in All of the Sciences *329*
8.5.1 Extending the VIM for the Human Sciences *329*
8.5.2 A Sociocognitive Caution About Properties in the Human Sciences *331*
8.5.3 In Human Sciences, Applied Instrument-Scale Development Comes Before Unit Definition *331*
8.6 Properties: Extensive and Intensive Metadomains *332*
8.6.1 Extensive *332*
8.6.2 Intensive *333*
8.6.3 Critique of the Extensive and Intensive Metadomains *333*
8.7 Ostensive Metadomain *334*
8.7.1 Ostensive Definition *334*
8.7.2 Ostensive Performance Demonstration *335*
8.7.3 Assessing Criterion Performance *Ostensively* at a Capstone Level or a Competition *335*
8.7.4 Ostensive Demonstrations Entail Reorganized Teaching and Assessment Roles *338*
8.7.5 Outline for a Possible Ostensive Demonstration Metadomain for Human Sciences *338*
8.8 Summary and Concluding Thoughts *339*
8.8.1 Lack of International Metrology Citations in Human Sciences Measurement Work *339*
8.8.2 Looking Ahead: We Will Not Sit Around Waiting *340*
8.8.3 Open Skills Network and Rich Skills Descriptions *340*
8.8.4 Putting Rich Skills Sharing "in Context," a Comparison to the History of SI Units *341*
8.8.5 As a Send-Off *342*
Acknowledgments *342*
8.9 Study Questions *343*
References *345*

9 Epilogue: Aspects of This Psychometrics Book *347*
9.1 Improbable and Unusual *347*
9.1.1 Unusual—The Metrological Slant This Book Takes *348*
9.2 Unusual Scientific and Pedagogical Perspectives *348*
9.2.1 An Unusual Scientific Perspective *348*
9.2.2 Unusual Content and Structure *349*
9.2.3 Mathematical Appendices 1–5 *349*
9.2.4 Software Computational Guides for Practice in Using Psychometric Tools *350*
9.2.5 Benefits of the Long Preparation for the Book *351*
9.3 The Value of Design and Domain Theory *351*
9.3.1 Domain-Specific Analysis and Modeling: Their Use in Assessment *351*
9.3.2 A New Metadomain Using Ostensive Definition and Ostensive Demonstration *352*
9.4 New Approaches to Measurement, Validation, and Educational Evaluation *353*
9.4.1 A Progressive Approach to Measurement *353*

9.4.2 Future Perspectives on Validity and Validation *353*
9.4.3 The Western Governors University Experiment and the Ostensive Demonstration of Skills *355*
9.5 Summary *357*
References *358*

1 Computational Guide 1 Classical Test Theory: A Model-Based Approach *359*
1 Classical Test Theory and the True Score Model *360*
1.1 Reliability with Multiple Test Components *360*
1.2 True Score Models for Multiple Items *361*
1.3 Latent Variable Analysis for True Score Models *361*
1.4 True Score Models as Factor Models *363*
1.5 Latent Variable Modeling of a Composite Score *365*
1.6 Inter-Item and Item-Total Correlations *366*
2 Estimating True Score Models as Latent Variable Models *368*
2.1 The Congeneric Model *368*
2.2 Tau-Equivalent and Parallel Models *370*
2.3 A Strictly Congeneric Model *371*
2.4 Item Reliability *371*
2.5 The Congeneric Model with Equally Reliable Indicators *372*
3 Composite Reliability *374*
3.1 Coefficient Alpha *375*
3.2 Bootstrapping Coefficient Alpha *376*
3.3 The Reliability Index for Congeneric Data *376*
3.4 Alternate Estimation of Coefficient Alpha *378*
3.5 Coefficient Omega *379*
3.6 Bootstrapping Omega *380*
3.7 Jointly Estimating and Comparing Alpha and Omega *380*
3.8 Omega If an Item Is Deleted *382*
3.9 Coefficient Rho *383*
3.10 Maximal Reliability and Optimally Weighted Indicators *384*
4 Standardized Reliability *385*
4.1 Standardized Alpha *386*
4.2 Standardized Omega *387*
4.3 Standardized Rho *389*
4.4 Standardized Maximal Reliability (Coefficient *H*) *390*
5 Summary *391*
References *397*

2 Computational Guide 2 Item Response Theory *399*
1 Item Response Theory (IRT) Models for Dichotomous Items *400*
2 Logistic (Logit) Item Response Theory Models *400*
2.1 The Rasch Model *401*
2.2 The One-Parameter Logistic (1PL) Model *401*
2.3 The Two-Parameter Logistic (2PL) Model *402*
3 Data Example *403*

3.1 Analyzing the Compressed Data *403*
3.2 Expanding the Compact Data and Analyzing the Expanded Data *404*
3.3 Estimating the Rasch Model *405*
3.4 Item Response Theory (IRT) and Categorical Item Factor Analysis (CIFA) for Dichotomous Items (1PL) *406*
4 Categorical Item Factor Analysis and Item Response Theory Models for Dichotomous Items *407*
4.1 The Rasch Model *410*
4.2 Estimated Ability Scores with the Rasch Model *411*
4.3 Relative Model Fit of the 1PL and 2PL Models *412*
4.4 Item Characteristic Curves (ICCs) and Item Information Functions (IIFs) *412*
5 Normal Ogive (Probit) IRT Models *414*
6 IRT Models for Dichotomous Items Using Mplus *415*
6.1 The Rasch Model *415*
6.2 The 1PL Model *416*
6.3 The 2PL Model *417*
6.4 IRT Parameterization for the 1PL and Logistic Rasch Models *417*
7 IRT Models for Ordered Categorical (Polytomous) Items *417*
7.1 Category Comparisons *417*
7.2 Model Specification *418*
7.3 Categorical Item Factor Analysis (CIFA) for Ordered Polytomous Categorical Items *418*
7.4 Model Equivalence, Nesting, and Comparisons *419*
8 Four Ordinal Polytomous IRT Models *421*
8.1 Model A: Item Discrimination Parameters and Step Thresholds *421*
8.2 Model B: Common Discrimination Parameter and Step Thresholds *421*
8.3 Model C: Item Discrimination Parameters with Item Location and Common Threshold Offsets *422*
8.4 Model D: Common Discrimination Parameter with Item Location and Common Threshold Offsets *422*
9 Software and Example Analyses *422*
10 Cumulative Probability Models Using Stata *423*
10.1 Model A: The Graded Response Model (GRM) *423*
10.2 Model B: The Graded Response Model with Common Discrimination (GRM-C) *424*
11 Adjacent Category Models Using Stata *424*
11.1 Model A: The Generalized Partial Credit Model (GPCM) *424*
11.2 Model B: The Partial Credit Model (PCM) *425*
11.3 Model C: The Generalized Rating Scale Model (GRSM) *425*
11.4 Model D: The Rating Scale Model (RSM) *425*
11.5 Summary *425*
12 Cumulative Probability Models Using Mplus *426*
12.1 Model A: The Graded Response Model (GRM) *426*
12.2 Model B: The Graded Response Model with Common Discrimination *427*
12.3 Model C: The Modified Graded Response Model (MGRM) *427*
12.4 Model D: The Modified Graded Response Model with Common Threshold Offsets (MGRM-C) *428*
13 Adjacent Category Models with Mplus *429*

13.1 Model A: The Generalized Partial Credit Model (GPCM) *429*
13.2 Model B: The Partial Credit Model (PCM) *430*
13.3 Model C: The Generalized Rating Scale Model (GRSM) *431*
13.4 Model D *431*
14 Model Implementation *432*
14.1 Cumulative Probability Models *432*
14.2 Adjacent Category Models *432*
15 Model Fit for the Example Data *433*
A. Appendix *434*
References *435*

3 Computational Guide 3 Factor Analysis *437*
1 Unidimensional and Multidimensional Factor Analysis *438*
1.1 Factor Analysis with Multiple Factors *438*
1.2 Exploratory and Confirmatory Factor Analysis *439*
1.3 The Independent Clusters Correlated Factors CFA Model *440*
1.4 The Correlated Factors Exploratory Factor Analysis (EFA) Model *441*
1.5 Higher-Order Factor Models *441*
1.6 Bifactor Models *442*
1.7 Data Example *444*
2 Correlated Factors Model *445*
2.1 The Correlated Factors Exploratory Factor Analysis (EFA) Model *445*
2.2 Communalities and Exploratory Structural Equation Modeling in Mplus *446*
2.3 The Correlated Factors Confirmatory Factor Analysis (CFA) Model *448*
3 Higher-Order Factor Models *448*
3.1 The Higher-Order Factor CFA *448*
3.2 The Higher Order CFA Model with Direct Indicators of the Higher Order Factor *450*
3.3 The Higher Order EFA/ESEM Model with Direct Indicators of the Higher Order Factor *451*
4 Bifactor Models *452*
4.1 The Symmetric or Complete Bifactor CFA *452*
4.2 The Bifactor CFA with a Reference Factor *453*
4.3 The Bifactor CFA Model with Direct Indicators of the General Factor *454*
4.4 The Orthogonal Bifactor EFA/ESEM with Direct Indicators of the General Factor *455*
4.5 Model Fit of Multidimensional Factor Analysis Models *456*
5 The Relationship Between the Higher Order and Bifactor Models *457*
5.1 The Higher Order Model as a Constrained Bifactor Model *457*
5.2 Loadings of the Indicators on the Higher-Order Factor Using the Higher Order Model *458*
5.3 The Bifactor Model as an Extended Higher Order Factor Model *461*
6 Reliability Estimation for Multidimensional Constructs *462*
6.1 Stratified Alpha *462*
6.2 Reliability Index with the Correlated Factors Model *464*
6.3 Reliability Index for the Higher Order Factor Model *466*
6.4 Reliability Estimation Via the Reliability Index in the Bifactor CFA *467*
6.5 Reliability Estimation for Higher-Order Factor Models Using Analytic Formulas *469*
7 Omega Reliability Estimation Via Reliability Formulas for Hierarchical Data *471*
7.1 Omega Hierarchical *471*

7.2 Omega Total *471*
7.3 Omega Hierarchical Subscale *472*
7.4 Omega Subscale *472*
7.5 Calculating Omega Reliability Coefficients from Factor Loadings and Residual Variances *472*
8 Expected Common Variance *475*
8.1 Expected Common Variance for the General Factor *475*
8.2 Expected Common Variance for Subscales *475*
8.3 Item Expected Common Variance *475*
9 Other Reliability Related Information *477*
9.1 Construct Replicability Index *477*
9.2 Average Relative Parameter Bias *477*
9.3 Factor Determinacy *478*
10 Special Situations in Reliability Assessment *480*
10.1 Negative Factor Loadings on the Specific Factors *480*
10.2 Correlated Indicator Residuals *481*
References *482*
Online Datasets and Services *483*

Appendix 1: Central Tendency: Mean, Median, and Mode *485*
Appendix 2: Calculation of Means, Variances, and Correlation Coefficients by the Algebra of Expectations *491*
Appendix 3: Matrix Multiplication *505*
Appendix 4: Bivariate Regression Analysis *519*
Appendix 5: Multiple Regression Analysis using Matrix Algebra and Stata *535*
References for the Five Mathematical Appendices *561*
Index *563*

About the Authors

Bruce L. Brown is Emeritus Professor of Psychology at Brigham Young University. He taught graduate and undergraduate courses from 1968 to 2024 in such topics as psychometrics, multivariate analysis, experimental design, psycholinguistics, and history of psychology. His research and writing have been in mathematical psychology, multivariate graphics, cognitive effects of infectious diseases, psychometrics, rapid reading, and acoustic indicators of emotion. He co-authored a 2012 Wiley book titled *Multivariate Analysis for the Biobehavioral and Social Sciences* with Suzanne B. Hendrix, Dawson W. Hedges, and Tim B. Smith. In July of 2025 he was appointed as one of several distinguished research advisors at Pentara Corporation, Mill Creek, Utah, a company specializing in clinical trial design for neurodegenerative diseases.

Joseph A. Olsen, Associate Research Professor in the College of Family, Home, and Social Sciences at Brigham Young University and director of the college's Research Support Center, has taught courses in intermediate and advanced statistical methods from 1994 to 2024. He has advised and assisted faculty and graduate students in a wide variety of research design and analysis projects. His areas of research interest include measurement, structural equation modeling, multilevel modeling, and longitudinal data analysis.

C. Victor Bunderson is currently a writer, Emeritus Professor, and entrepreneur retired from Universities and University Councils and from offices and boards of computer-assisted education companies. He served as Assistant and Associate Professor and Director of the Computer-Assisted Instruction Lab at University of Texas at Austin, then as Professor and Director of the Institute for Computer Uses in Education (ICUE) at Brigham Young University, overlapping in 1998 with advisory councils at Western Governors University that culminated in the WGU Assessment Council. He retired from this council in 2024 after 26 years. WGU is an on-line, high-enrollment, competency-based International University with headquarters at Salt Lake City, UT. In this book it is used as an example of reorganizing conventional Universities and colleges to become student-centered using psychometrically sound competency-based assessments and mentoring for student success. He is deeply indebted to his own mentors at the University of Utah, University of Southern California, and at Princeton as a Psychometric Fellow, then his mentors/colleagues at Texas and BYU. A special time was at ETS again in 1996–2001 as VP of Research Management. He learned from many remarkable colleagues. Finally he owes so much for understanding gained from the dynamic and creative people at WGU and fellow assessment council members about using measurement to promote notable educational change.

Foreword

Psychometrics, Test Theory, and the Latent Factors Model by Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen is a notable volume that unfolds the concept of measurement in a way that reveals one of its fundamental properties, indeed one of its virtues, that is seldom appreciated by those in the field of psychology who rely most heavily on measurement. It is an aspect of measurement that may be most appreciated and best understood by psychologists deeply engaged in the theoretical and philosophical grounding of the discipline. Expressed in the simplest terms, this approach to measurement and data analysis frees both of them from the metaphysical and epistemological overreach that is the discipline's legacy from positivism. The importance of this freedom for psychological science and practice is difficult to overstate.

Expressed in nonmathematical terms, the genius of the approach articulated in this volume begins with the recognition and affirmation that there is a logic built into numbers, as there is in every language, such that they can convey meaning, consistency, and relationality embedded in the human world of action and purpose. The language of numbers can, therefore, reflect the consistency, relationality, and meaning in the world even at a level that might be very difficult for people simply being-in-the-world to articulate. It is not necessary to invent an "unconscious mind" to account for this phenomenon, as a positivist theoretical approach might require. Rather we can safely assume that there is always more to any "saying" than what is explicitly "said." That this all occurs within the language of numbers is one of the most important and psychologically relevant revelations in this book.

The principle articulated here opens up new methodological as well as metaphysical perspectives and projects for the field. Positivism has essentially a single use for numbers and measurement, i.e. to support metaphysical claims by validating the existence of some concept developed for the purpose of explaining some phenomenon in classical deterministic terms. In contrast to this is the possibility, illuminated in this volume, that the language of numbers, i.e. data, can enhance understanding and provide evidence regarding the content, and meaning of behavior as lived experience, which might otherwise be opaque to practitioners of positivist social science and even to human beings themselves. In other words, rather than simply verifying metaphysical claims in the positivist tradition, a legitimate use of the serious and substantive statistical analysis is to open up possibilities for investigation and understanding from within the lived world. This project is reflected in the book in terms of the discovery of "latent variables". In nonmathematical laymen's terms, the innovative power of measurement and analysis can articulate the inarticulate but deeply meaningful world of human life as lived in context, giving the deeply contextual human world of lived experience meaningful mathematical expression. As is clear throughout the various sections of this volume, and nowhere more than in the technical mathematical chapters, this is no mean feat.

The chapters in this volume present for the reader the mathematics and concepts underlying the project articulated in this short review. Thus, this volume, admirably, brings sophisticated mathematical principles to bear on an issue of fundamental importance in our understanding of our own humanity. This is a worthy project indeed.

Richard N. Williams, PhD
Professor of Psychology
Brigham Young University
Provo, UT 84602

Preface

This book honors Petr Blahuš, deceased scholar and scientist of note, but it is not a Festschrift written by former students and colleagues. Instead, it honors his work and his memory by placing connected parts of his collection of valuable manuscripts within a broader framework of relevant work since his death. It reshapes some of his prescient but unfinished ideas to become possibly very useful contributions to contemporary psychometrics. Through this broader framework we, as continuing co-authors, are striving to make factor analytic methods, IRT, and related tools more accessible. These have great utility, not only in the development of good measuring instruments, but also in connecting them better to theory—an attainment most earnestly pursued by Petr in his manuscript. To be useful as a textbook, it must also have the easy-to-understand explanations and conceptual, mathematical, and computer tools that enable students to become proficient in the use of psychometric and statistical tools. These have been added both within the chapters and as mathematical appendices and computational demonstrations and guides.

The three co-authors became Petr's students in the sense of our thorough study of the manuscript he entrusted to BYU Professor Bruce Brown in 2010. We have been taught by his fresh and unique insights into the science of psychometrics. For several years we have been focused both upon understanding the depth of his approach and insights, and how they fit with the many psychometric advances of the past two decades. This book has taken a long time to reconstruct, reconsider, and write, and is published in 2025 as a textbook for Educational, Psychological, and Kinanthropometrical measurement, perhaps also useful for other measurement-oriented disciplines in the human sciences.

In 2014, Victor Bunderson had agreed to write a Foreword, which by definition is written by someone other than an author. That task required reading the evolving chapters and appendices of the final manuscript which immersed him into reading the drafts that were available. In 2018 his time became freer and he began to write the explanatory text and updating text surrounding Petr's sections in several of the chapters. Later he began to write what became Chapter 8, the Metrology Chapter. We as co-authors have written this Preface that includes an Introduction to how the book is conceptualized and structured. It shares our gratitude for those individuals and works most influential to this synthesis of Petr Blahuš's manuscript with later work, including the additional references we cite, and our view of where the topics of Psychometrics, Test Theory, and The Latent Factors Model are going.

How this book came to be. Petr Blahuš was a world-travelling scholar. He was a visiting professor in Paris, Athens, Slovenia, Croatia, Slovakia, and the United States. His many publications were in English, Russian, French, and Czech. His longest visit was in 2001–2002 for a full year at Brigham Young University, Provo, Utah, United States. It was there that he met and worked with the

colleagues who would later become his successor authors of a book he envisioned as a capstone to his work. He drafted some sections while he was at BYU, and was granted permission by Professor Erin Bigler, with whom he worked closely, to use as data examples some of Erin's studies using the CERAD Battery. You will see these in the text. Early manuscript material was in use with his students in Prague several years before 2008.

Unfortunately, Professor Blahuš became ill and passed away in 2010 at the age of 66. Shortly before his death Bruce Brown and his wife Susan visited him in Prague. Dr. Brown was and is a Professor of Psychology and Quantitative Methods. He and Petr had become close friends when they worked together at BYU in 2001. During the Browns' 2010 visit to Prague, Petr gave Bruce the manuscript, and Bruce promised to make every effort to see it completed and published. This book came to be only because Bruce and his colleagues believed it was a strong manuscript and well worth the time and dedication it would take to complete it. It was also necessary to bring topics up to date to integrate the influence of important new developments in the field, and to add student-friendly materials and study aids for a textbook format.

The original manuscript had a proposed title of *Structural Modeling,* and a provisional subtitle for use with students: *A Methodology-Based Introduction to Behavioral Statistics, Test Theory, and the Latent Factors Model*. The title for the completed and published book is *Psychometrics, Test Theory, and the Latent Factors Model*.

Psychometrics: We follow his methodological approach and explain psychometric methods through the many analyses that a psychometrician performs, but especially in establishing *scales* and their *validation*.

Test Theory: He wrote of both classical test theory and also modern test theory, that is, Item Response Theory—recognizing it has flowered and evolved during the past 15 years—perhaps more than Petr expected.

Latent Factors Model: This part of the title implies the need for a particular type of powerful measurement scales and the logical and mathematical processes necessary for their Validation. *Valid Measurement Scales* becomes, therefore, one of the four overarching topics of this book without appearing in the title.

More completely, the four main topics are: psychometrics, *valid measurement scales*, test theory, and the Latent Factors Model. At the time he was trying the manuscript out with students, he was serving under the title of Petr Blahuš, PhD, DrSc, Professor, Division of Methodology, Department of Kinanthropology; and also Humanities, Faculty of Physical Education and Sport, Charles University, Prague.

Kinanthropometrics. When we speak of Psychometrics, we also embrace Petr's particular focus, Kinanthropometrics. Petr was methodologically oriented. His methodologies came from these disciplines and from mathematical statistics. Petr's department of Kinanthropology was proposed and created as a separate department at Charles University, Prague, Czech Republic, in 1989. It offered a distinctly scientific approach to higher studies in physical education, sports science, therapy, and sports medicine, and replaced earlier doctoral degrees in those fields. A specialty within that field is Kinanthropometry. It provides specialized metrology and methods for sports science and physical education. He gives examples from these areas in his manuscript. However, the manuscript is not a treatise on the specialty of Kinanthropometry. It falls squarely under the more general term of Psychometrics, with the subtlety of mental measurement strengthened by the contrast to objective physical performance.

It seems apparent from his body of work that the study of human movement is more readily grounded in reality than is explaining invisible processes of thinking and feeling. The sequence

of physical movements in a complex performance are often observable. This may be partly the reason he repeatedly demands clear and explicitly investigated and understood constructs—in non-movement domains involving invisible cognitive, conative, or affective states as well as in the more spread-out in time and visible sequences of physical movements. This is a unique perspective. Sport increases one's focus on realism because the skilled movements of the human body in sports science are also accompanied by repeated replication and thus to improvement over practice trials. Since comparability is vital in documenting new records and placement in events, sports venues have become well standardized. Norms at various ages and levels of practice gradually become documented and spread downward linking through international competition the athletic educational systems of each nation.

Petr Blahuš believed in good science in the behavioral domains where he worked. He believed that such science required mathematical models that were as accurate as possible in depicting an underlying reality. He resisted operationist pretenses that the operations could stand for the underlying reality. He was a strong methodologist but sought methodologies that he had traced to their historical origins; ones whose practical advantages and weaknesses were understood. He looked for insights from appropriate philosophies of science.

His early academic career focused on methodologies, factor analysis in particular, and their foundations in science and mathematics. He took additional studies in the philosophy of science and added to those studies throughout his life. After his dissertation in 1974, he rose within eight years to become Acting Head of the Division of Mathematical Methods and Experimental Techniques within the Division of Psychology, Czechoslovak Academy of Sciences. After another eight years there (simultaneously serving as Associate Professor at Charles University, Prague), he became Vice Dean for Research at Charles and later served as Vice Rector of Charles University for three years while remaining active in publishing during those years of full administrative responsibilities.

Attaining the rank of Full Professor in 1992, his inauguration lecture and defense was on *the measurement and modeling of motor abilities in developing their scientific theory*. It was an extension of his dissertation work. As we will see, measurement and modeling of abilities is inseparable from understanding a progressive theory of what is going on at varying levels of ability, so that the meaning of the measurements can be interpreted. Thus, contact with the qualitative world is improved. He fully understood how measurement levels and structure need theory both to guide measurement and to provide meaning for interpretation and use. *Interpretation and use* are one chamber in the double-chambered beating heart of contemporary issues in the validity theory of assessments. The other chamber is a full understanding of what measurement, and indeed, all of metrology is about.

Content and Structure of this Book

To understand this book, a reader will come to understand its principal author, Petr Blahuš, and what he discovered, stood for, spoke for, and sought to pass on before his death in 2010. This book displays the results of how the three co-authors who completed it over the past seven or eight years have put his "book within a book" into the physical and conceptual chapter structure you now see. It was at first not so much a book as a sequence of brief sections dealing with specific psychometric topics. The sections of Petr's manuscript, mostly his words lightly edited but occasionally largely rewritten, comprise the "book within a book." The reader will encounter these sections, numbered in the sequence which he followed. This sequence is independent of the chapter numbering. His manuscript had 135 smallish sections, highly cross-refenced (which at least partly explains the long

gestation period in producing the book). Of these we were able to use 70, while having to leave out several additional very good materials on such topics as philosophy of science and neuroscience from this already large enough book.

Each of Professor Blahuš's sections in the book is designated with the symbol "**§n**" at the beginning of each **n**th section from his manuscript as they are organized within Chapters 1 to 7. We retain this pattern of **§**s for consistency and cross-referencing among the first seven chapters, the ones based upon his work. We drop this pattern of section numbering and revert to the commonly used "double numeration" style in Chapters 8 and 9, created primarily by Victor Bunderson.

If there is not a **§** [1] symbol in a heading in Chapters 1-7, it is introductory or explanatory material added beyond the original manuscript. Chapter introductions are found under the first section in each chapter: 1.1 for Chapter 1; 2.1 for Chapter 2; and so on. There are also chapter major subdivisions: for example 3.1, 3.2, 3.3, and 3.4 in Chapter 3. These also have an introductory and explanatory function. Finally, there is a summary subdivision at the end of each chapter, followed by study questions for students, and ending with the references following the summary and the study questions.

In reading the book within a book, the reader should be aware that Petr Blahuš was not only strong on methodology, and on finding the founder(s) and foundation(s) of a concept, but also on mathematical representation. Multilingual Petr wanted these two quotes to appear at the beginning of the book:

Immanuel Kant:
Ich behaupte aber, daß in jeder besonderen Naturlehre nur so viel eigentliche Wissenschaft angetroffen könne, als darin Mathematik anzutreffen ist.

I only insist that, in any fundamental study of nature,
the genuine science is dealt with only as far as mathematics is included.

Leonardo da Vinci:
Nessuna umana investigazione si puo demandara vera scienzia s'essa non passa per le matematiche dimostrazione.

No human investigation can be called genuine science
if it is not accompanied by mathematical proof.

Because his mathematical explanations are not always easy reading, Bruce Brown and Joe Olsen, who were teaching at BYU, early saw the need for the addition of Mathematical Appendices and Computational Guides for statistical analysis software. Bruce wrote, with the intent of being student-friendly, with example problems involving small sample sizes, such that students could easily work out the math. The explanations used multiple representational forms for concepts—written explanations, equations, the datapoints, the graphical representations, and the matrix algebra representations. He tested these and drafts of some of the chapters with students. Joe Olsen made accessible to students the latest software, available in most universities, which enables work with larger datasets. He created Computational Demonstrations as supplemental materials that students could work through with the expectation that some of their professors might also provide additional data for them to practice with using the code.

1 This symbol (**§**) is a glyph. There are many kinds of glyphs. Wikipedia defines a glyph as "any kind of purposeful mark." This particular glyph is a section marker, frequently used when citing sections of legal code. We follow Prof Blahus's example in this "book within a book" by using them to mark and cross-reference the 70 sections in Chapters 1 through 7. A few glyph-based sections, especially in Chapters 3–5, were not originally Petr's but were assigned **§**'s. They are recognizable in that they refer to him, or in other ways denote it is not his voice.

Development of This Book

We have seen it as a priority to put his manuscript concepts in their updated and expanded forms into the contemporary context of what has been happening to the book's main topics since 2010. The subsequent writing effort had no funding beyond what the individual authors provided. Bruce Brown and Joseph Olsen had full teaching and research roles at BYU and limited time to turn a long manuscript into a useful book on the intended subjects. Until 2018, Bruce and Joseph were negotiating with the publisher, organizing it into chapters, and beginning the writing of appendices and computational guides that would give it enough scaffolding to be accessible to their students. Victor Bunderson left BYU teaching in July 2002, but continued advising doctoral students, and conducting research at his EduMetrics Institute while learning and advising about on-line, competency-based assessment through service on the Assessment Council of Western Governors University. He felt a kinship to his long-term BYU colleagues in quantitative methods and with Petr Blahuš from a meeting in 2002 at a seminar where Petr had invited his good friend Roderick McDonald. Ultimately, a kinship developed with an unseen author as well as the visible ones, through the struggles of turning a worthwhile manuscript into a useful contemporary textbook.

A Dialectic with Our Senior Author

The team of continuing authors have worked well together and have imagined themselves in a dialectic not only together, but with Petr through his written words, his references, and the memories of discussions with him in the past. Our respect for the senior author is strong, and we generally have consensus with him about the balance of theory and psychometric/experimental investigation, science philosophy, and the importance of research on latent variables and factors. Several such issues in such a dialectic are mentioned in the chapters, but here we focus on how sports science, one of Petr's favorite topics, especially the Olympic Events every four years both summer and winter, can serve as an example for the general science of psychometrics.

The Olympic events offer spectacular human demonstrations of incredible mastery. They are infused with national pride. They honor fandom in many sports of interest—shared by media to the far corners of the world. The highest level of mastery of the Olympic subdomains, or disciplines, is demonstrated publicly within the games. There, the very best athletes, backed by the very best individual coaches, work to understand and master specific and inspiring feats in standardized venues. It is anchored by the claim that the measurements that take place are comparable over times and locations. The new Olympic world records demand comparability, as do the fine distinctions between gold, silver, and bronze medalists. These strongly require a trust in the accuracy, comparability, and fairness in the conditions surrounding the measurement events. The immense practice time and years of qualifying tournaments make the human and related science models (relevant to how to succeed) of great interest to all the levels of performance—at younger ages and at ages leading up to the spectacle of the Olympic Games.

Insights into the Human Testing Enterprise Through Comparison with Olympic Testing

Let us begin with the design, development, and use of the "measurement instruments" for games that have come to be accepted by every nation, (albeit with some grumbling about unfair refereeing or biased ratings in artistic or team sports). Shown in italics below (with arrows,→, on the left of

each line) is a high-level model of sequential flow in the creation and hosting of winter or summer games as part of a four-year Olympiad. When in the book we replace this with the analogous sequence for development and use of an educational or psychological test battery the similarity will be apparent. There are lessons to be learned from the Olympics and its stress on measuring the very best in the world in each game with instruments designed to be comparable over time and place. Some of the essential tasks and requirements in preparation for each Olympiad are:

- ➔ *Select domains for measurement and specify the number and capacity of venues to house the measurement instruments for each game a venue will support.*
- ➔ *Design and develop the measuring instruments* that meet established standards while fitting well within existing and newly constructed venues within the host cities.
- ➔ *Plan the schedule of the total Olympic Games and the logistics* for transportation, housing, and feeding of the contestants and for managing the expected audiences.
- ➔ *Establish scales familiar to the public, or readily understandable, and calibrate each instrument* to those established scales with an understanding of the high levels of attainment the scales must measure.
- ➔ *Host the events and provide all personnel and facilities* to assure the measurements are properly taken, the rules are followed, and fairness prevails.

This flow of tasks and events, from conception through design and development and leading to a spectacular implementation with world-wide interest, suggests many comparable points for consideration in the practice of psychometrics.

Most Olympic Events Use Units from the International System of Units (SI)

An important theme in this book is the importance of International Metrology and its expression in Scientific, Applied, and Legal concepts and organizations. Choosing metric system units of the SI to measure the shortest times, highest jumps, heaviest weights lifted, longest throws is natural, as nations in these days of the reborn Olympic games have already agreed to use these units. As Chapter 8 shows, International Metrology also has much to offer the psychological and educational testing industries. In these areas many graduates in psychometrics-related fields may find their careers. What are some of the lessons we can learn from the International Olympic Games that can be applied to psychometrics and indeed to the more general field of International Metrology?

1) ***Domain Analysis:*** *First-class measuring instruments are developed for narrow and well-specified domains.* The meta-domain of Olympic Sports and Artistic Festivals must be subdivided first into the type of sport, then into specialized domains, called disciplines, in order to develop measurement instruments within performance venues that include scoring systems. The instruments and scoring systems have similarities, but each one is unique. This book features, especially in Chapters 4 and 5, a critique of measurement that does not fully recognize that domain-specificity is fundamental to good measurement.
2) ***Models & Theories:*** *Top athletes are accompanied by superior coaches who understand the models and theories that enable powerful coaching feedback. They understand the strengths and weaknesses of their mentees*. They are also good at motivation and stress management. Peak performance is worth the time and money to get the best models of top performance supplied by great coaches who help their athletes win the medals. Models and model-based reasoning,

including models of the human athlete, the task, and other matters, enable coaches to guide improvement following each qualifying run, and after each permitted replication of a performance. These topics are addressed in most chapters, but Chapters 3, 4, 5, and 6 introduce validity as influenced by a new definition of measurement for all the sciences. The implications of this definition for bringing human sciences into the world of international metrology are addressed mainly in Chapter 8 on metrology. Human performance models differ at varying levels of attainment, unlike physical, chemical, and engineering Metrological units, which are uniform at every level.

3) ***Instrument Development:*** *Positive values and consequences justify the costs of instrument development*. First-class measuring methods and instruments have led to the world-wide success of the Olympiads. In Chapter 3 on the theory of the validity of test scores, definitions of validity which make test developers responsible for unexpected negative consequences have been a point of contention. Despite this, *Fairness* continues to gain more space in each succeeding publication of the *Standards for Educational and Psychological Testing*. In the Olympics, the positive values and the consequences in the lives of the champions and all who participate are not arguable, nor is the importance of fairness. If positive consequences are not strong in a domain of human activity, there may not be the necessary social interest and consequent economic feasibility for investing in instrument development.
4) ***Scale Standardization:*** *There are detailed protocols to assure standardized methods and fairness for each event in each sports subdomain or discipline*. These protocols use referees to disqualify participants who violate the rules of each game. Prior to participation, there are now blood tests to rule out doping.

These four points and the flow of preparing for and executing an Olympic Testing Event has its counterpart in Chapters 3 and 4 to describe similar processes of psychometric and educational instrument development and use. To summarize, *domain analysis* and consideration of the potential positive consequences and values to the population of people involved in the domain is the first stage of instrument development. This is followed by creating *models and theories* for each sport. *Developing the instrument* is next, including defining its scale and the range of the scale needed for current and future records. The final phase is *standardization*, the development of careful management plans for the performance demonstrations by the athletes in each venue, and in contact with each scaled instrument. It is "a game-day protocol" with standardization for fairness and with specified action in case of rule violations. Clearly, fairness for all and equal treatment must be guarded at all costs to retain enthusiastic international cooperation and support.

We hope that this book will lead many readers to envision at a broader level, not only through the example of sports science, but also through the disciplines of international metrology a bigger picture for the future of our field. Perhaps psychometricians can learn from the example of sports disciplines at the highest level of true criterion scores how our disciplines, or domains, can be improved as well. We need to achieve validity, comparability, fairness, and to keep improving the entire process of designing, developing, testing before use, then bringing on measurement instruments and events that are enticing to our audiences.

Acknowledgements

This book would never have reached this point without the other two co-authors, Joseph Olsen (who has been my trusted tutor in using statistical and psychometric software packages for several decades) and Victor Bunderson (who knows more than anyone with whom I am acquainted about the developing science and practice of psychometrics). I invited each of them to join as co-authors, and fortunately each of them accepted.—Bruce L. Brown

As related in the Preface, we have appreciated very much the writings of our senior author that have guided us in discovering a refreshing perspective regarding "the measure of all things human." It has been a difficult journey, but rewarding, illuminating, and often inspirational. We also owe a debt of gratitude to Kimberly Monroe-Hill, our Senior Managing Editor for her patience and attentive diligence. Many thanks also to Deenadayalu Govindanagaraj, our Content Refinement Specialist, for wisely and kindly training us in conveying the final revisions of the page proofs in an efficient manner.

We also acknowledge that we would not have accomplished our goals for the production schedule without the assistance received from our friends and colleagues at Pentara Corporation, Mill Creek, UT. In particular we thank Suzanne Hendrix (CEO), Patrick O'Keefe (Director of Statistics), Tyler Duke (Statistician), Pablo Zubeldia (Statistician), Allison Brownlee (Proof-reading and Edits). We also appreciate several computer rescues from Eric Hendrix (Systems Administrator).

C. Victor Bunderson—This book goes against convention in stressing International Metrology as the proper context for Psychometrics. It clarifies unneeded disputes about Validity, and emphasizes Design Science, and Ostensive Demonstrations of Competency. Some deep thanks is due to friends who helped us authors directly or though their published work in each area.

International Metrology: Chapter 8 has its own acknowledgements about who gave us the metrology understanding, and it started with physicist and former director of the Voltage lab at the National Institute for Standards and Technology, Dr. Yi-Hua Tang, who became my mentor in Thailand in 2019, and later reviewed the chapter. In 2022, we discovered Mari, Wilson, and Maul's incredible 2021 book on Measurement Across The Sciences, and it brought to life Petr's tables and figures where he argued for increasing scientifically sound bridges between the empirical world and the world of interpretation and use. Mari, Wilson, and Maul gracefully named that the "information world" and showed how three intentionally designed processes through an instrument could perform direct measurement. Mark Wilson and William Fisher, who joined Mark at Berkeley later, had previously helped me and my Edumetrics colleague Dr. Van Newby, and graduate students publish our self-funded work done through the Edumetrics Institute. Dr. Fisher had shared with me some drafts and later published more on the economic and social impact of metrology.

Validity was unsettled scientifically and theoretically in 2009 when Petr cited Robert Lissitz's book on *The Concept of Validity*, but the metrology spectrum of Scientific, Applied, and Legal allows us to see that conflict is not necessary. Perspective of where we are on those three aspects of a full metrology system, along with the concept of an "information world" clarifies it. On the scientific side we have to credit Borsboom and his colleagues and their definition of validity. Mari, Wilson, and Maul do that, but it leaves validity only in the scientific world as mere accuracy. In the "information world" the instrument's values must be interpreted, used, and evaluated for fairness and reliability. This is where Messick, Kane, and many others centered the definition of validity—in the domain of applied and legal metrologists. They did not mention measurement. They did not mention instruments. These omissions are good for legal insularity, but not for validity. Of direct impact was my relationship with Sam Messick during the years as Vice President of Research Management while he was Vice President of Research. I became familiar with some of his writings then, and used his writing extensively with students after returning to teaching and research at BYU in the early 1990s. I owe the decision to do that to Dr, Brian Gong, a former student and then colleague at ETS. He told to go back and work with some graduate students and work it through in detail. I did, wanting to find out what Messick meant by "Domain Theory". I am so grateful for this advice and also that he helped me directly in writing on consequential aspects of validity. Bob Mislevy's work while I served at ETS and since has been very central. We ended up using his definition of validity from his wonderful 2018 book on *Socio-Cognitive Foundations of Measurement*. Howard Wainer and coauthors on *IRT and Computer Adaptive Testing*, and other advice from Howard gave me a textbook and we authors a dust cover for this book—the high-jumping analogy to adaptive IRT came from him.

Design Science. I was a closet design scientist from my first reading of Herbert Simon's first book on the Sciences of the Artificial, or design science, while a student at Princeton. Design research under Edumetrics was initially centered in achieving Validity. Later, Evidence-Centered Design had been defined by Bob Mislevy with numerous colleagues who have made it very widespread and understandable. This now functions as a Design Language which permits others to write improvements for the Domains they analyze and the models they synthesize. David Merrill induced me to BYU from Texas by offering design theory by which we could justify the features we put into the TICCIT system. Moreover, he offered staffing using many student workers to recruit from BYU's extensive media programs. In addition to Dave Merrill, there was Ed Schneider, Gerald Faust, Andy Gibbons, Charlie Reigeluth, Greg Kearsley, Fred O'Neal, Tom Zane and many others.

Ostensive Demonstrations of Competency. A student can turn around and become your mentor, and Robert Mendenhall, a PhD student I advised at BYU in Instructional Psychology and Technology, became the first President of Western Governors University and appointed me to the assessment council, where I learned first-hand what competency-based learning systems can accomplish, organized with people in ways foreign to conventional higher education. Others added greatly, including current president Scott Pulsipher, Sarah Demark, Peter Ewell, Donna Sundre, Randy Bennett. Brian Gong. WGU also provides a new source of information for William Fisher and others who tout the impact of metrology on social and economic outcomes. In Mendenhall's dissertation and in his practices as President of WGU, he also specified competency exams for the entire University. Scott Pulsipher, the current president, continues these evaluations and uses the Gallup organization to evaluate WGU programs and graduates success. The Epilogue cites those data. It shows the value to the economy of these students, many from minority, rural, or first degree of anyone in their family. It provides more validation for William Fisher's claims about the impact of metrological practices on social and economic outcomes, and for these authors enthusiasm for framing psychometrics in international metrology.

Outside of these categories, as Bob Mendenhall says, I persuaded him to forsake law school and try educational technology. Two professors at University of Utah convinced me enter scientific psychology instead of law school toward other careers. They both did so by mentoring to me as a research assistant. Lyle Bourne had me conduct the student subjects through sessions with his remarkable electronic machines that presented concept learning stimuli, and measured responses and response latencies. This foreshadowed future computer data presentation and collection with scientific controls and model based reasoning. He also gave me a co-publication in the *Journal of Experimental Psychology* that we wrote together. Calvin Taylor involved me in factor-analytic studies of creativity of NASA scientists. He was the first of three students of LL Thurstone that mentored me. The other two were at Princeton as a psychometric fellow: Harold Gulliksen and Ledyard Tucker, who consulted from Illinois with the fellows.

I owe the most for learning how to set up a laboratory at a university, mentor doctoral and masters students to work in it, get it funded, and advertise it with good publications, from Psychologist and Dean of Education, Wayne Holzman, who recruited me out of Princeton to the University of Texas at Austin to set up the Laboratory of Computer-Assisted Instruction that he chaired. He has been a wonderful mentor and friend for many years before he passed in 2019.

Joseph A. Olsen—I've been honored to work with Bruce and Victor on this project and would like to acknowledge the contributions of the many discussions with them and with two additional and particularly valued colleagues at Brigham Young University, Scott A. Baldwin and Richard R. Sudweeks. Finally and especially, I'd like to acknowledge the insight and support of my own dear and true brother, James B. Olsen, the real practicing educational psychometrician in the family.

Bruce L. Brown—I was fortunate to have received strong career preparation at McGill University primarily under the guidance of Wallace E. Lambert (dissertation chair and my mentor in psycholinguistics and dissertation chair) and Donald O. Hebb (department chair, "physiologizer" and neuroscience pioneer—ranked as the 19th most cited psychologist of the 20th century). McGill was an apprenticeship-based program with only two classes, *Hebb's Seminar* and a statistics course from George A. Ferguson.

There were about 10 of us in the doctoral class of 1966. The seminar began with the question "Are you a monist or a dualist?" Nine answered "monist." As the last one to answer I said "dualist" and everyone laughed. Dr. Hebb reminded them that these issues are not settled and we should be open to listen to the case for each person's position. The next week he gave us a list of theorists who had a major impact upon mainstream scientific psychology. Each person on the list was a prominent behaviorist. I had the temerity to suggest that there may be other defensible positions besides behaviorism. I still remember his measured and wise response, "All right, Mr. Brown, you may have one-half hour in class next week to defend that position."

I went to the library and buried myself in the work of each of my favorite psychologists from my undergraduate studies but each possibility evaporated when I considered defending it. I agreed to present Clark Hull's work. A few weeks later, Dr Lambert handed me a copy of Noam Chomsky's small monograph *Syntactic Structures*. Within 20 minutes I reached pages 19–21 on Markov processes. I was hooked, and wishing I had seen it a few weeks earlier, in time for Hebb's Seminar. A few months later Noam Chomsky himself came to Montreal to work with Professor Lambert and I was invited to a private debate between Dr. Hebb and Dr. Chomsky at the home of Dr. Lambert. It solidified my commitment to cognitive psychology and psycholinguistics.

In 1968 I was hired by BYU, my Alma Mater as an ABD "Instructor." I was drawn to quantitative methods, and during the summer of '69 I embarked on another life adventure, the NSF-Funded University of Michigan Summer Fellowship in Mathematical Psychology. As our text we were given

a pre-publication manuscript copy of a new book by Clyde Coombs, Robyn Dawes, and Amos Tversky, *Mathematical Psychology: An Elementary Introduction*. The course was taught by the three of them and nearly a dozen of their colleagues including David Krantz (who along with Amos Tversky) was one of the four authors of the three *Foundations of Measurement* volumes, referred to in Chapter 1 of this book. For many years I have been particularly drawn to Tversky's work both in these three volumes and also in Tversky and Kahneman's prospect theory which appears in Figure 2.8 and the accompanying footnote 7 of Chapter 2. Prospect theory was awarded the Nobel Prize, but it was presented in 2002 to only Kahneman because Tversky had died 6 ½ years earlier in 1996 at the age of 59.

Dr. William J. Strong (PhD from M.I.T.) acoustics professor, Dr. Alvin C. Rencher (PhD from North Carolina) a multivariate statistician, and I received an NIMH grant in 1970 to study the acoustic causes of personality impressions. I am grateful for all that I learned from each of them and for the works we published together. Dr. Rencher and I also worked extensively together in 1994 when I became his reader as he prepared his 1995 Wiley book *Methods of Multivariate Analysis*[1] for publication. It was a great learning experience and led to my own 2012 Wiley book (which I sometimes call "Rencher lite") titled *Multivariate Analysis in the Biobehavioral Sciences,* which I wrote with three co-authors: Suzanne Hendrix, Dawson Hedges, and Tim Smith. Whereas Rencher's book was written for graduate students in statistics, ours was written using a "simplest case" approach to make it accessible to motivated but non-mathematical undergraduate psychology students. We used it successfully for 13 years to teach undergraduate students at BYU.

In his 1995 multivariate book, the first edition, Dr. Rencher commented that factor analysis (which grew out of the work of Charles Spearman) in contrast to principal component analysis (which grew out of the work of Karl Pearson) is considered by many statisticians to be controversial. Many statisticians still share that view, as I did for many years. When they do have a need for factor analysis they will use the principal components method of factor analysis because it is true to Pearson's deterministic mathematics and also "accounts for more variance." However, when Dr. Rencher created his third edition he was joined by a newly minted young PhD, William F. Christensen, who had learned in graduate school that the non-deterministic approach is superior because of the mathematical power of the "latent factors model." The rationale of the superiority of the latent factors model is explained very clearly by Petr Blahuš in the last section of Chapter 7 of this book: "§70 Dimensionality by the LC Factor Model vs. PCA 'Factor Analysis'." The logic underlying the latent factors model can be obtained from a careful reading of Chapter 6, and in particular actually trying the "Tapping Test" which is explained in §61 of Chapter 6. The test can actually be performed and then analyzed as directed in Computational Demonstration Study Questions, CD7 and CD8. It is the best demonstration I've seen of how the "latent common factors model" works. Petr Blahuš created it for his students quite a few years ago.

For much of my career I still had a strong preference for the principal component method of factor analysis. This becomes very obvious when you read what we wrote on pages 200 to 222 of Chapter 4 in our 2012 book:

Brown, B. L., Hendrix, S. B., Hedges, D. W., and Smith, T. B. (2012). *Multivariate analysis for the bio-behavioral and social sciences: A graphical approach*. Hoboken: John Wiley & Sons.

1 I had used a variety of multivariate books in my graduate course in the years before Rencher's first edition was published in 1995. I used it in my graduate multivariate analysis class as soon as it was published, and it worked substantially better than any previous book. His 1995 book was a masterpiece of comprehensiveness, clarity and completeness. His third edition is also, but with the addition of the latent variables material that he and Dr. William Christensen have added.

The principal component method of factor analysis can be very useful for some things, but not for establishing a latent variable. With its geometrically deterministic mathematics, it is very good for creating useful multivariate graphs (as shown in Chapters 4 and 5 of the Brown, Hendrix, Hedges, & Smith book) that can display higher dimensional spaces more economically in smaller space projections, but it can't be used to establish a latent variable as advocated by Borsboom and his colleagues (2004) in about the tenth paragraph of Section §25b of Chapter 3. Latent variables are established by demonstrating "local uncorrelatedness," as an indicator of the principle of local independence, as explained in Section §33c of Chapter 3.

Joseph Olsen, a co-author of this book, who understands the logic of latent variables better than anyone I know, tried to explain this crucial difference between principal components and latent variables to me over thirty years ago, but I didn't get it. Petr Blahuš also tried to explain it to me when I first met him years ago, but I still didn't get it, but now, after nearly a decade of working on this book with Petr's clear explanations it is finally very clear to me and I can finally see the great value in being able to discover and demonstrate the existence of a valid latent variable.

About the Companion Website

This book is accompanied by a companion website:

www.wiley.com/go/psychometrics

The website includes Study Questions and Answers for each appendix.

1

The Psychometric Quest

CHAPTER MENU

1.1 Quantification in Psychometrics: Historical Origins, 3
- 1.1.1 (§1) Why Are Mathematical Methods Central to the Empirical Sciences? Why Are Statistical Methods Used in the Behavioral Sciences?, 3
- 1.1.2 (§2) The Specific Mission of Statistical Methods in the Behavioral Sciences, 4
- 1.1.3 (§3) Why Do We Use *Behavioral Statistics* and Not Just a Branch of Applied Mathematics?, 4
- 1.1.4 (§4) Some Remarks on the Historical Relationships Between Psychology and Statistics: The Development of the Correlation Coefficient and Least Squares Regression, 5
  - 1.1.4.1 The Correlation Coefficient and Least Squares Regression, 6
  - 1.1.4.2 Correlation, Regression, and Heritability, 7

1.2 Variables, Scales, and Data, 9
- 1.2.1 (§5) *On Terminology I*: Attributes with Categories, Variables with Values, Tests with Scores, and What About "Data"?, 9
  - 1.2.1.1 Variables and Data, 9
  - 1.2.1.2 Scientific Rigor, Representation, and Stochastic Modeling, 11
- 1.2.2 (§6) Measurement Scales and the Representational Theory of Measurement (RTM), 12
  - 1.2.2.1 The Ferguson Committee, 12
  - 1.2.2.2 Stevens' Representational Theory of Measurement (RTM) and of Scale Types, 13
  - 1.2.2.3 Stevens' Four Measurement Scales, 13
  - 1.2.2.4 Foundations of Measurement, the "Received View" of RTM, 16
  - 1.2.2.5 The Contrast Between Foundations of Measurement and the Science of Applied Psychometrics, 19
  - 1.2.2.6 Metrology and Measurement Units, 21
  - 1.2.2.7 Measurability in the Behavioral and Social Sciences, 23
- 1.2.3 (§7) *Scientific* and *Stochastic* Variables: When May "Data" Be Treated By Statistical and Test Theory Methods?, 24

1.3 Statistical Methods: Location, Dispersion, and Statistical Inference, 25
- 1.3.1 (§8) Group Level of an Attribute—The Mean Location of the Group's Scores on an Interval Scale, 25
- 1.3.2 (§9) Comparing Median vs. Mean: Which One Is "Fair"?, 28
- 1.3.3 (§10) A Psychometric Look at the Mean $\bar{X}$, Item "Difficulty," and p, 28
- 1.3.4 (§11) Intra-group and Inter-individual Variability—*Dispersion* of Scores, 31
- 1.3.5 (§12) A Small Numerical Example, Finally, 33
- 1.3.6 (§13) Does the Larger Head Circumference in the Sample of Sons Than Fathers Hold for the Whole Population? A Preliminary Sidestep to Statistical Inference, 35
  - 1.3.6.1 An Estimate of the Probable Error of a Point Estimate, 38
  - 1.3.6.2 Classical Probability Theory and the Binomial Distribution: Creating a Simple Stochastic Model, 40
  - 1.3.6.3 The Binomial Distribution as a Simple and Clear Example of a Sampling Distribution of Proportions, 46
  - 1.3.6.4 Elaboration of Higher Level Calculations for the Binomial Distribution, 49

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

1.3.6.5 Multinomial Demonstrations of Two Fundamental Principles of Statistical Inference, 52
1.4 Summary, 56
1.5 Study Questions, 56
References, 62

A friend and colleague of ours is a practicing psychometrician at a company specializing in leadership measurement. A few years ago, in preparation for a report in elementary school, his son asked him about his occupation. He told him he was a "psychometrician." He practiced the pronunciation of the word with him and told him the kinds of things he did. He explained that his life's work was focused on finding ways to measure important but subtle things, that is, things that are difficult to measure. That evening he asked his son how the report went. It became clear that even after practicing the word "psychometrician," he could not recall it. The best the boy could muster was to report to the class that his father was "a psycho," and that he tries to find ways to measure things that are hard to measure. His report to the family generated a good laugh with no threat to his father's professional confidence.

However, in the 1930s, there was a major threat to the integrity of psychometrics as a discipline generally, and indeed to all the behavioral and social sciences, as described in Section §6. The story is there related of the Ferguson Committee of the British Association for the Advancement of Science (physics vs. psychology) in their ongoing debate regarding the status of measurement in the social and behavioral sciences. The debate continued for years, throughout most of the 1930s. The vote of half of the committee was that it is not possible in principle to measure psychological attributes scientifically. There is much at stake here, for it calls into question whether any one of the behavioral sciences could truly be considered to be a science. Fortunately, several psychometric and statistical scholars since that time (Luce & Tukey, 1964; Luce, Krantz, Suppes, & Tversky, 1990; Krantz, Luce, Suppes, & Tversky, 1971; Suppes, Krantz, Luce, & Tversky, 1989) have successfully refuted that challenge, with assistance from recent advances in metrology (Stock, Davis, de Mirandes, & Milton, 2019). There is also the pragmatic argument. Psychometric theory and methods have, over the past century, developed into a necessary and useful branch of applied science in a variety of areas including psychological, educational, and medical testing.

In 2012, Ian Stewart (2012) published a book about "seventeen equations that changed the world." We would like to propose an eighteenth equation—the analytical method known as factoring. Without factoring, there would be no psychometrics as we know it. Factoring is the foundational method for the development of psychometric theory and practice. It began early in the 20th century with the publication of Karl Pearson's (1901) paper that introduced what later came to be called the "principal component" analytical method. A few years later, Charles Spearman's (1904b) paper introduced the parallel method that would become "factor analysis," an expanded application of the mathematics of principal component analysis to address psychometric questions. Spearman's purpose in developing the factor analytic method was to examine the elemental structure of human intelligence.

In the intervening century, factor analysis has become the essential method for examining not just intelligence but many other constructs that cannot be directly observed. It enables psychometricians to examine whether test items reliably cluster together reflecting a common underlying psychological construct. For example, do such things as knowledge of state capitals, or of vocabulary words, and the rapidity with which one can solve a puzzle inter-correlate in a way that reflects an underlying factor that could be called intelligence? Does intelligence consist of one overarching component that best describes what it means to be cognitively capable, or are there a variety of types of intelligence such as judicial wisdom, creative and mechanical skill, emotional intelligence, social facility, and interpersonal effectiveness? Perhaps intelligence is better understood as a constellation

of related but distinct neurocognitive parts such as working memory, processing speed, and verbal reasoning.

What about personality tests? Are they capable of predicting how people will respond in situations? Can they be used to guide therapeutic interventions, or to clarify a clinical diagnosis? Factor analysis has proven to be effective in addressing psychometric questions such as these in a wide variety of areas. Without factor analysis, the past century of development in psychometric theory and methods would not have been possible.

1.1 Quantification in Psychometrics: Historical Origins

1.1.1 (§1) Why Are Mathematical Methods Central to the Empirical Sciences? Why Are Statistical Methods Used in the Behavioral Sciences?

This book represents the culmination of an academic quest of nearly 50 years by the primary author, Dr. Petr Blahuš, brought to completion after his death with assistance from several of his friends. Central to his quest has been a belief in the essential unity of all scientific endeavor. In his own words:

> I should introduce myself as an advocate of the belief that there exists a general unity of scientific knowledge. This monistic personal philosophy of mine stresses that *in principle*, there exists an *essential unity of scientific method* in the empirical sciences. This philosophical view has nothing to do with positivism, and especially not with its two extreme versions: *scientism*, which claims that there is an absolute identity in the methods used, and *reductionism*, which moreover states that the identity is also found in the object under study (i.e. psychology can be finally reduced to the principles of physics). Therefore, my personal philosophy could be perhaps called a relative and moderate methodological monism.

Petr maintained that this essential unity permits us, step by step, to walk on the endless path of learning, to improve our limited knowledge and to find connections among the many and varied branches of science. The pieces of our knowledge, though insufficient and temporary, must be organized or *systemized* so that we can find and recognize certain regularities. In this systemizing effort, certain auxiliary devices are not only helpful but necessary. These auxiliary devices are the *formal methods* of the abstract sciences, including mathematics, logic, and a variety of graphical devices such as the diagrams of molecular structure. These formal methods are essential to the pursuit of empirical science. Complex and subtle relationships among empirical observations cannot be uncovered and described without them. Imagine trying to do chemistry without structural diagrams of molecules or atoms. Or try to imagine a world in which the science of physics must be pursued without the availability of calculus.

The myriad branches of the empirical sciences are found at various stages of progress in their use of these formal methods. In past centuries, most applications of formal methods came primarily from quantitative and deterministic mathematics. They were also closely related to specific empirical problems in the established fundamental sciences, such as classical physics and astronomy. Later, and even up to the present, methods adapted to qualitative concepts and able to deal with a certain degree of uncertainty were employed in the human sciences such as psychology. Because these methods are not so closely related to specific research applications, they have broader and more flexible applicability. Statistical methods have become the prototype. They are typically the beginning of a formalistic infiltration into a branch of the human empirical sciences, that is, behavioral or social sciences such as psychology, kinesiology, sociology, political science, economics, organizational behavior, and education.

1.1.2 (§2) The Specific Mission of Statistical Methods in the Behavioral Sciences

The role of statistics, as with all kinds of formal methods, must be determined with regard to the overall purpose of the particular science. In short, the main purpose of all sciences is to approach an *understanding* of nature and to disclose the *laws* by which nature is governed, as far as possible due to the partiality which is allowed to us. In the empirical sciences, that purpose is sometimes pursued by a general methodology that includes the two fundamental scientific functions of *explanation* and *prediction*. The explanation is typically based on *general scientific laws*, or even *universal laws*, which might even have a *causal* character deeply rooted in the nature of the observed phenomena. On the other hand, prediction[1] often may be based on certain superficial regularities. These *partial and descriptive superficial empirical regularities* can offer empirical estimates, quite often directly for some practical use and possibly without any deeper explanation.

As we can see, all these procedures are based upon a formalized *systemization*. Thus, we could characterize the role *of statistical methods* in the social and behavioral sciences as that of *contributing to our knowledge of scientific laws and regularities* through formal symbolization. This is carried out on two levels. One is the *exploratory* role in discovering new regularities. The second one is *confirmatory*, that is, testing the theoretically derived hypotheses with a special aim not just to confirm or reject but to specify a *form of the lawfulness* under scope. Here the "form" should not be understood too narrowly, say as a curve relating two quantitative variables to one another. The specification of form can be accomplished in a wide variety of ways, including verbal description, graphics and diagrams, and mathematical expressions.

Design is a complementary and necessary third component to the main two functions, that is, to the *exploration* that leads to hypotheses, and to the *confirmation or disconfirmation* of proposed explanations. To obtain a confirmation of a hypothesized theoretical proposition, a meaningful explanation, requires both careful instrument design and careful experimental design to set up the confirmation situation and rule out alternative explanations. Thus, we might say the processes of *exploration* and *explanation* require the *design* to attain reproducible scientific merit.

In this way, statistical methods, as well as other formal methods, have become part of a gargantuan network of many approaches to systemizing the relationships and links between observable data and scientific theory. Gradually, the formal apparatus and the original empirical core of the branches of science are becoming amalgamated and mutually synthesized. In this process, the separate and occasional applications are transformed into several foci of true interdisciplinary scientific disciplines. In these focal synthesized areas, the "empirical-formal tandem" works in cooperation that allows each of the components to influence the other and to initiate mutual development in a creative way.

1.1.3 (§3) Why Do We Use *Behavioral Statistics* and Not Just a Branch of Applied Mathematics?

This is a reasonable question, for "statistics" is being used as an abbreviation for "mathematical statistics," which didn't become a fully recognized mathematical discipline with axiomatic background until the end of the 19th century and the beginning of the 20th century. One aspect of the answer has been indicated in the last section: some branches of science require an adjustment or

1 We could use the term prediction in a broader sense, almost in a metaphorical sense of "the guessing of the unknown." This may give a connotation of "forecasting" that would be misleading. A better term would be *estimation*, which would cover all three time-horizons: (i) *contemporary estimation* as of diagnosing a psychotic syndrome from certain symptoms, (ii) prediction in its narrower sense of forecasting future success in school from admission tests, or (iii) retrodiction of a certain prenatal brain injury based upon symptoms found in adulthood.

adaptation in the mathematical procedures to accommodate an uncertainty component in prediction and toleration for a broader variety of approaches to quantification. Also, there are several relatively stable synthesized interdisciplinary areas that psychology and statistical science have in common. You find them, for instance, in psychophysics in the form of theories and models for the measurement of sensory stimuli. They also are found in psychometric methods such as classical test theory (CTT) models and item response theory models, or latent structure models of intelligence.

Another aspect of the fit between the behavioral sciences and the discipline of statistics can be illustrated in the historical development and mutual influence of the two disciplines. Their relationship is really unique and could be compared to the historical relationship between classical physics and deterministic mathematics including geometry. In some social and behavioral sciences, the use of statistics started in terms of isolated applications which helped further the development of empirical science. However, with psychology, the situation was quite different. Historically, psychologists themselves tried to create statistical methods for solving their specific research problems and this became an impetus for the consequent mathematical treatment of the problem on a more formally sophisticated basis. This can be illustrated with the example of the mutual interconnection between methods of correlation and theories of intelligence. Without the pressure and demand that psychological questions imposed upon statistics and mathematics, the development of methods such as correlation, regression analysis, factor analysis, and related latent-variable developments would have certainly been delayed, if not precluded.

Furthermore, the synthesis of psychology with statistics is not limited to meeting the general aims of scientific inquiry. Psychology also has very practical clinical branches, which make it parallel to similar situations in medicine with regard to biological science. In these areas, the intention is to produce practical benefits through clinical treatment. But every treatment must be preceded by diagnosis, and for that reason, psychometrics has become an essential tool in many kinds of clinical work. A neuropsychologist who specializes in the diagnostic evaluation of nervous system disorders is a good example of this. In a sense, "psychometrics" is an interdisciplinary theory of behavioral diagnostics, a synthesis of psychology, psychobiology, psychiatric medicine, and statistics. From another, and perhaps an oversimplified, point of view, psychometrics could be described as the theory of measurement errors in behavioral evaluation technology. The psychometric theory of psychological tests and questionnaire inventories, dealing as it does with the problems of validity, reliability, and measurement error, can be introduced as a model illustration of the genuine interdisciplinary synthesis of an empirical "soft" social science with the formal "hard" sciences of quantification.

1.1.4 (§4) Some Remarks on the Historical Relationships Between Psychology and Statistics: The Development of the Correlation Coefficient and Least Squares Regression

The earliest and most fundamental contributions to the science of psychometrics date to the second half and end of the 19th century, and the beginning of the 20th century. Under the influence of Darwin's theory of evolution, the mutual relationship between the influence of heredity and environment began to be studied. There were two main lines of study and two great men who accomplished major advances in genetics in the last half of the 19th century. The first was Gregor Mendel, a monk who recorded the effects of hybridization on qualitative attributes (like color) of beans and peas which he grew in the monastery garden in Brno (a Czech town in the Bohemian Kingdom, in those times a part of the Austrian Empire). More important for psychology, however, was the second man—an Englishman by the name of Francis Galton—whose prodigious and varied scientific

labors were partially enabled by being born into wealth. His early accomplishments were in geography, but, influenced by his half-cousin Charles Darwin, his attention turned to human biology, anthropology, and intelligence, with a particular interest in the problem of "genius" and its hereditary roots (Galton 1869, 1889). In contrast to Mendel's qualitative approach, Galton focused on quantitative, measurable attributes and used them to compare inter-individual differences among consecutive human generations, with special attention to individual "deviations" from the population mean. Many years later, the famed statistician Sir Ronald A. Fisher explained the mathematical inter-relationship between the Galtonian and Mendelian approaches, in a classic paper containing the first use of the term "variance" (Fisher, 1918).

1.1.4.1 The Correlation Coefficient and Least Squares Regression

Influenced by phrenological assumptions about the relationship between abilities and the shape of the skull, Galton attempted to evaluate the degree of heredity by the similarity of skulls between fathers and sons. As well, he invented tests to measure several psychological abilities, especially what he called "natural (mental) ability," a predecessor of IQ, which, he believed, was primarily innate. He tried to find mutual cross-generational similarities as indicated by corresponding father and son deviations above or below their generation's average. For this purpose, he invented a quantitative index of statistical similarity. Later, with mathematician, philosopher, and his close junior friend, Karl Pearson, they jointly improved the index to be between 0 and 1 in absolute magnitude and called it the coefficient of "co-relation" (Galton, 1888).

This index is well known today as the Pearson product-moment *correlation coefficient* (Pearson, 1896), or the PPMCC (see Appendix 2). The name was derived from the formal similarity of its computational formula to the mathematical formulas for mechanical moments[2] in classical physics. Mathematically, the correlation coefficient is based upon the "least squares" method.

Much earlier, in the late 18th and early 19th centuries, two other men formulated and applied the least squares mathematical method. It would later become least squares regression analysis (see Appendix 4), which in its current form is closely related to the correlation coefficient. The first of these two men was Carl Gauss, who in his astronomical observations of the orbits of planets in 1794 or 1795 worked on the problem of *errors of observation*. He tried to fit and smooth the imperfectly observed measurements in a mathematically optimal way. Thus, while the exact *true* orbits remained unknown, they were best approximated by the fitted values, with optimal fitting achieved by minimizing the sum of squared deviations of the observed values from the fitted values. The second man was Adrien-Marie Legendre (1805), who published the first paper on the topic.[3] Although the discovery of this "least squares" method preceded the development of the correlation coefficient by nearly a century, in modern treatments of the two topics, the correlation coefficient is typically introduced first, and then the computational method for least squares regression is taught as an extension of the logic and mathematics of the correlation coefficient, as shown by Appendix 4 (regression) being preceded by Appendix 2 (correlation) in this book.

---

2 Regarding the correlation coefficient as an application of mechanical moment calculations from physics Petr Blahuš, our senior author quipped, "So, do not worry if some time you meet a strict statistician who tends to stay on the ground of pure mathematics as far as possible and explains the correlation coefficient as 'simply a central standardized second-order mixed moment,' and then adds, 'perhaps you could use it somehow in psychological research, too'."

3 Gauss is credited by many with being the discoverer of the least-squares method. However, some would argue that Legendre should be accorded that honor, since he published the first paper on the topic. See Plackett (1972) for a weighing of the arguments and evidence in favor of each side. See also Stigler (1981) for additional information that has come to light, and Brown, Hendrix, Hedges, and Smith (2012, p. 373) for a two-paragraph summary of this major priority dispute in the history of statistics.

1.1.4.2 Correlation, Regression, and Heritability

The explicit formulation of the correlation coefficient expressing the degree of association between various pairs of human attributes, x and y, was motivated and pursued by Galton, and then mathematically formulated by Pearson. The degree of association between fathers and sons was interpreted as the amount of their common genetic background—the heritability—with regard to a certain attribute. The next step was to focus on the problem of which attributes are more heritable and which less. Intelligence, which he referred to as "mental ability" should have been mostly innate, according to Galton's theory, perhaps more than other attributes. Therefore, it was of particular interest to compare the correlation, that is, the degree of heritability, in this attribute. Galton devised the first scientific mental measurement tests, making him the founder of psychometrics. He also created a wide variety of "anthropometric" measurements of the chief characteristics of the human body, which could also be tested for heritability in comparison with the heritability of mental characteristics.

Thousands of subjects were tested with Galton's anthropometric and mental measurement methods at health exhibitions, and in his laboratory at the South Kensington Museum. Many were willing to pay a small fee for the privilege of participating and receiving a copy of their results. He gathered so much data that it was not analyzed until well into the 20th century. Galton's methods were somewhat consonant with the phrenological speculations of the times, that is, the idea that the form and size of the skull had a direct relationship to "mental faculties." For example, a higher father-son correlation between skull circumferences than between their body heights might support the hypothesis of a higher heritability of "mental ability" than of physical body characteristics. The Gaussian least-squares fitted line enabled the estimation or prediction of a son's head circumference from that of his father.

Galton's interest was in developing methods for the prediction of mental capabilities from one generation to the next. The relationship was thought to be decomposable into a *predicted* part, which he hoped would be due to *heredity*, and an *unpredictable* part, which he presumed would be due to the *environment*. This was an *important step in the general methodology of science*. It was the beginning of the turn from a *deterministic approach* to measurement, as in physical science, which had allowed only the researcher's errors of observation as a kind of methodological (epistemological) imperfection of science, to a new incorporation of the *stochastic*[4] *conception of the world*, allowing some *ontological uncertainty* (that is, actual uncertainty in the makeup of the laws of the universe) to be an essential component of nature and its laws.

Of course, the controversy of epistemological vs. ontological uncertainty was well known as a philosophical problem well before the time of Galton. In some well-established sciences, such as physics, it was discussed much earlier. For instance, Laplace (1814/1952) articulated the view that uncertainty is at the root epistemological. He proposed that if we were able to "know all the forces that set nature in motion and all positions of all items of which nature is composed" then we would have the future before our eyes. On the other side, and more recently, Peirce (1901/1957) articulated the ontological view. He proposed that uncertainty is at least partly inherent in Mother Nature—a property of natural law, including the laws of physics. Galton's achievement was to bring a stochastic emphasis into the life sciences and the behavioral sciences. In this, he stood on the same side as Peirce.

The fitted estimation line, which was intended to make a graphical estimate from fathers' to sons' values of head circumferences, was called by Galton a "reversion line." The estimate could also be produced arithmetically by a simple equation that describes the line. The *estimation equation* as a

4 A process that is stochastic could also be referred to as "probabilistic," randomly determined, or at least somewhat attributable to random chance as opposed to determinism.

device for producing the numerical prediction of sons' from fathers' head measurement values is shown here using fictional illustrative data, numerically assumed to be in inches.

$$(\textit{predicted}\ \text{son's head circumference}) = 1.02\ (\text{father's head circumference}) + 0.04$$

$$\hat{Y} = 1.02X + 0.04$$

In general, the equation for (linear) statistical estimation contains two coefficients a, and b, with the resulting estimated value y marked by a hat ^, which signifies that it is an estimated value.

$$\hat{Y} = bX + a$$

The numerical values of the above coefficients can be interpreted as follows:

$a = 0.04$ means that there is a tendency for the generation of sons to have a slightly larger head circumference than the fathers, a 0.04-inch shift,

$b = 1.02$ tells us that a father who has a 1-inch larger skull than another father should have a son whose skull circumference is, on average, about 1.02 inches larger than the son of the other father.

Moreover, coefficient b also means that the inter-individual spread among sons is larger than the inter-individual spread in the fathers' generation.

Galton's observation that the predicted values $\hat{y}$ for the sons of fathers with particularly high or low deviations from the group mean tended to be closer to the mean of the new generation was called "regression toward the mean."[5] He interpreted this phenomenon as an effect of, and evidence for, Darwin's theory regarding the influence of natural selection. The name stuck. In modern statistical practice, the best-fitting linear estimation equation is now referred to as the *regression equation*, and its geometric representation is the *regression line*.

In his 14 June 1920 lecture to the Royal Institution, Pearson himself informs us that the symbol r that we now use for the Pearson product moment correlation coefficient also comes to us from Galton: "Here for the first time appears a numerical measure r of what is termed 'reversion' and which Galton later termed 'regression'." This r is the source of our symbol for the correlation coefficient, which was really the letter of "reversion" not of "regression" (Pearson, 1920, p. 33; see also Piovani, 2008). In this same paper, Pearson corrects his earlier error in crediting Bravais (1846) and his product sum with the discovery of the correlation coefficient and accords that honor to Galton.

The concept of correlation, as well as the Pearson product-moment correlation coefficient as a numerical index, is now essential and ubiquitous. It has become indispensable in practice as a data-analytic tool, but it is also central and vital as a theoretical concept. As the absolute value of r approaches 1, the accuracy of the regression estimation grows to its maximum. If the accuracy reaches a sufficient level, the equation can be used for *practical estimation*, that is, for prediction. This descriptive statistic, r, enables us to evaluate the predictive validity of psychological diagnostic methods, thus providing a rational foundation for using a test to predict the "fit" of an applicant for a certain job or perhaps using a test to predict how well a student will do in graduate school.

Correlation and regression are the fundamental underlying concepts of the modern psychometric theory of tests. They make it possible to establish the reliability and the possible diagnostic error of a newly constructed test battery or questionnaire. They are also central to a host of other diagnostic quantities such as dimensionality, construct validity, and consistency. At a more theoretical level,

5 This statistical phenomenon, oft-observed, can be stated simply: if an observation of a random variable is extreme on first measurement, it will generally be closer to the mean on the next measurement.

they constitute the building blocks for the wide variety of factor analytic methods, as well as the path-analytic structural equation models and the associated latent variables that enable one to identify theoretical scientific concepts, referred to as *constructs*.

1.2 Variables, Scales, and Data

In the terminology of behavioral research methods, there are several areas that are affected by semantic ambiguity and confusion. The problem has deep methodological roots, and creates several areas of confusion, with important and troublesome consequences. We will clarify several of the terminological issues in Section §5 that follows.

Following the introduction of these clarification issues, Sections §6 and §7 give a brief historical account and discussion of a central question for psychometrics, "Is measurement in the social and behavioral sciences even possible?" and the corollary question of whether these disciplines, if they do not have bona fide measurement, could appropriately be considered sciences.

1.2.1 (§5) *On Terminology I*: Attributes with Categories, Variables with Values, Tests with Scores, and What About "Data"?

The term "score" is usually used narrowly to refer to the results of "tests" in the sense of psychological tests, such as sensorimotor tests, admission tests, standardized questionnaires, and so forth. However, in the behavioral, educational, and social sciences generally, *Test Theory* in its broad meaning is conceptualized as a *general theo*ry of the *diagnostic quality* and *standardization* of *any* method used to measure or classify behavior, whether in empirical research or in clinical or educational practice. Besides the term "scores" there are also other related terms, such as "values," "categories," "alternatives," or "levels," used to refer to types of behavioral classifications or measures.

1.2.1.1 Variables and Data

Another ambiguously used term is "variable," which takes on a wide range of somewhat differing meanings. For some people "a variable" is still joined with the idea of a mathematical variable, taking on many possible values, and being used in contrast or opposition to a mathematical "constant" which has a single value. Such a conception, which focuses on quantity, and distinguishes between a variable *quantity* and a constant *quantity*, is too narrow for a general methodology, due to its automatic assumption that we are dealing only with the well-defined numerical quantities of physics and other foundational empirical sciences. Moreover, it suggests the naïve and simplistic idea of the metric axis represented by real numbers, which is taught in elementary mathematics in school.

Is gender to be considered a variable? Is ethnicity a variable? Each obviously must be considered as such and have a place in the social or behavioral sciences if these sciences are to be based upon empirical observation. However, even at first glance, there is an obvious problem. There is an intuitive mismatch between, on the one hand, "*gender*" as an independent *variable X* predicting "*heart attack*" as a dependent *variable Y* (with both as empirical binary categories), and those simple independent and dependent numerical/mathematical variables X and Y which everybody knows from high school mathematics. The sense in which the term "variable" is used in these two settings is obviously quite different.

Speaking in a simplistic way, the mutual relationship between those two notions of "variable" is the topic studied by a *representational* theory of measurement (RTM) (Section §6), which deals with the problem of how well numbers *represent* or reflect empirical reality, and, on the other hand,

how well a real-world attribute may be represented by numbers, that is, by the formal ideal of a numerical scale. Therefore, it is first necessary to distinguish between two worlds, namely the *substantive world* of actual things, actual *entities* and their *attributes*, which stand as an original *reality* to be explained; versus the *formal world of symbols*, very often *numbers*, with its *syntactic* rules such as arithmetic operations, through which we try to represent the first world, the world of reality.

Thus, to explicate the notion of "variable," we must distinguish between (a) the *empirical attribute* itself, for example, "strength of attitude," which is thought of in its "natural," that is nonformalized original existence; and (b) the *formalized attribute,* as coded by symbols such as numbers. The primitive empirical attribute (a) could be provisionally organized by dividing it into *categories*. In this conception, we define the *category* as a *discrete possible alternative* outcome that an *unformalized attribute* may take on. However, in this sense, a category should *not* be assumed to be only purely qualitative (as in some statistical texts, where the purely qualitative "categorical data" are treated in a special way). Thus, even an attribute of a fully quantitative and continuous nature, like body weight, could be assumed and used as discrete. It could be used to sort people into categories given by some level of resolution. A subject's body weight numerically formalized as 140 lb. could mean it falls into the category represented by an interval between 139.5 and 140.5.

The formalized attributes (b), usually consist of symbols that can be easily processed. For example, in computerized processing perhaps ASCII characters would be preferred in general, and in particular, *numbers* are usually the most convenient. The formalization may occur as in the case of so-called keyed categories in an inventory with pre-printed response options, such as Likert scales, multiple-choice items, and so forth. Thus, the notion of a *mathematical variable quantity*—as known from high school mathematics—is just a very particular case of such a formalization. For some people, the sole act of formalization or symbolization is enough to allow the use of the term "variable" for any attribute, even if it is unclearly defined. Some are content to use terms like "values" or "data" to refer to the formal symbols, especially numbers, assigned to the categories in any ambiguous or even rather mysterious way. It is the task of measurement theory to clear up such ambiguities.

Another confusion may come from homonyms, such as when the word "*data*" is given several fundamentally *different and technically specific meanings* in different areas. In *computer processing*, for example, a programmer might refer to any information necessary for running a program as the "input data." For instance, the name of the data set, identification codes for names of subjects, etc., are also part of the input to the program. However, these are not genuine research data of scientific interest. Therefore, it is necessary to be careful in a discussion with specialists of different fields to distinguish between genuine research data and the "data," which constitute the input to a computer. Similarly, numbers on a sheet of paper might be assumed to be "the data" and automatically processed by a careless would-be statistician as if they were fully metric.

Another specific meaning of the term "data" is its use in latent variable modeling. The "data to be modeled," as input to a modeling procedure, need not be the directly observed or measured values but could be already-computed values, that is, values derived from the originally observed and coded ones. For someone skilled in latent variable modeling, for example, correlation coefficients can stand as the target "data" to be modeled or explained in terms of a latent common factors model such as one of the methods of factor analysis.

In addition, occasionally in *qualitative research* any laboratory records or field-research documentation, such as videotaped records of the behavior of children playing, or tape-recordings of narrative interviews, are sometimes called "data," in a kind of short-cut jargon. It might be wiser to refer to these kinds of preliminary observations as something like *raw research material* to avoid confusion. Some qualitative researchers advocate the definition of data as "everything that has been recorded" (Majerova & Meyer, 1999, pp. 135–141). We would suggest that perhaps the term "data"

should be reserved for the transformed state of such raw research material after it has been coded through a well-planned and logically defensible process. There is a contradictory aspect to referring to raw observations as data. Would we say that after obtaining data we then apply a coding procedure to the data? Would we then need to distinguish between the "uncoded data" and "coded data"? It seems somewhat unusual that "analyzing data" might result in their "coding." It is more usual to think of data as being ready for analysis *after* they have been coded.

What has been said about the importance of coding in qualitative research applies to every area of behavioral research. Data coding is a crucial step in the scientific process. Similar data coding, verification, and evaluation processes are also central to quantitative and even experimental research. If this foundational process is slighted, the later stages in the scientific process—data analysis, interpretation, and theoretical integration—are compromised. Haphazard or arbitrary data construction systems or strategies that are not well-reasoned may not fulfill the necessary conditions and requirements for measurement, and hence cannot be recognized and treated as "genuine data."

1.2.1.2 Scientific Rigor, Representation, and Stochastic Modeling

The question of whether various approaches to formalization may be recognized as producing "data," that is, bona fide research data that can support analysis within the framework of behavioral statistics and test theory, depends upon three broad preconditions: (i) scientific rigor, (ii) principles of representation/meaningfulness, and (iii) stochastic modeling.

The first area, scientific rigor, deals with general scientific foundations and can be considered in two parts: (ii) *standardization* of variables, and (ii) *theory-based anchoring* of variables. *Standardization* of the observational procedures includes several topics from classical measurement theory, such as establishing the reliability and validity of the variables and extending to other quality control methods that are essential to the art and science of testing. The explanation of these concepts will require additional preliminary knowledge. Hence, the discussion will resume later in the book. *Theory-based anchoring* of the variables has to do with so-called "construct validity," that is, establishing a variable in relation to other variables within the network of scientific lawfulness. It has to do with questions of scientific theory building and its corroboration by empirical data. These principles will also be dealt with in later sections, particularly in Chapter 3.

Whereas these two aspects of "scientific rigor" relate to the *practice* of psychometrics and will constitute a major part of what is taught in the remainder of the book; the remaining two preconditions—representation/meaningfulness and stochastic modeling—are both primarily *formal*. Representation and meaningfulness have to do with logical and philosophical foundations and will be dealt with in the discussion of representational theories of measurement (RTM) in Section §6.

Stochastic modeling is also formal. It is firmly based in statistical probability theory (discussed in Sections §7 and §13) and is closely associated with the concept of a random variable.[6] Traub (1997) identifies it as number two of the three major developments that made the emergence of psychometric theory possible. The other two are a recognition of the presence of errors in measurements and a concept and index of correlation. In Section §7, we will briefly explain why the decomposition of observed test scores into "true score" and "measurement error," and the recognition of measurement error as a stochastic random variable, is so important in establishing bonafide research variables. This theme will continue to appear in various places throughout the remainder of the book.

6 See Section §13, particularly the Section titled "Classical probability theory and the binomial distribution," for simple examples of stochastic modeling, which is one of the three preconditions for "bona fide research data" as introduced in this subsection of Section §5.

1.2.2 (§6) Measurement Scales and the Representational Theory of Measurement (RTM)

1.2.2.1 The Ferguson Committee

In 1932, the Ferguson Committee of the British Association for the Advancement of Science was appointed to consider and hopefully resolve the question, "Is it possible to measure sensory events, that is, human sensory experience?" Much is at stake in this question, for it is really the question of whether psychology or for that matter any of the behavioral or social sciences (sociology, political science, economics, etc.) can be considered to be sciences. The committee, consisting of delegates from physics (Section A of the association) and psychology (Section J), met for seven years with little progress. One committee member, complaining about the rigidity of opinions and the spirit of disagreement in the meetings, quipped that year after year, his colleagues "came out by that same door as they went in" (Stevens, 1946, p. 677).

In its last year of debate, the committee focused upon a particular psychophysical scale, the Sone scale of loudness (Stevens & Davis, 1938), which claimed to establish a mathematical relationship between an established physical measure of sound intensity, and a subjective scale of auditory sensations. As expected, in this final year the responses of the committee were again bifurcated and extreme, and again divided into the same two opposing camps. A statement from the negative camp is particularly illuminating, and on first look even compelling, arguing that "any law purporting to express a quantitative relation between sensation intensity and stimulus intensity is not merely false but is, in fact, meaningless unless and until a meaning can be given to the concept of addition as applied to sensation" (Final Report, p. 245).

The focus in this statement on "the concept of addition" as the central issue in whether a measurement method can be considered scientific mirrors N. R. Campbell's (1920, 1928) philosophy of measurement that held sway for over 40 years. Drawing upon the distinction between *extensive* measurement and *intensive* measurement, Campbell held that *fundamental* extensive measurement,[7] that is, acceptable scientific measurement, requires an empirical *concatenation* operation that corresponds to the numerical operation of addition, a goal beyond the reach of subjective measurement methods. The measurement of weight provides a simple example of what Campbell means by additivity in fundamental extensive measurement. Consider a beam balance with a 3-g weight in the left pan, tipped to the right by having a 5-g weight in the right pan. We then "concatenate," that is add, a 2-gram weight to the left pan and the weights are now in balance. The numerical addition operation of $3 + 2 = 5$ is *isomorphic* with the empirical concatenation operation. The two correspond to one another.

It could be said that the empirical concatenation of weights in a beam balance and the plus operation on real numbers form an isomorphism with one another. However, if this kind of concatenation operation is a *necessary* condition for establishing a valid scientific measurement scale it would follow that most psychological variables are not scientifically measurable. For example, it would be difficult to imagine what "empirical concatenation" could mean for IQ scores (are two persons with an IQ of 100 the equivalent of a person with an IQ of 200?), or for that matter, even for loudness judgments. The Section A members of the Ferguson Committee were asking the obvious next question. If those doing behavioral and social research cannot establish the

7 Fundamental measurement refers to measurement that is primary as are length and weight, not derived from other measurements as are volume and density. "Ostensive fundamental measurement" is measurement that is primary, but also concrete and demonstrable, as when for example we actually demonstrate measurement of length by laying down a yardstick repeatedly in measuring the length and width of a garden. The word "ostensive" denotes a way of defining by direct demonstration, that is, by pointing. The example of gram weights in a balance given in this paragraph is also an example of ostensive measurement.

scientific/rational basis for their *measurements* with a concatenation demonstration, how can they then use the scientific method, including measurement and experimentation, in their research?

1.2.2.2 Stevens' Representational Theory of Measurement (RTM) and of Scale Types

Stevens' response to the Ferguson Committee's evaluation of the Sone scale of loudness, and the general question of whether sensory events can be measured, appeared six years after the Ferguson report, in what was to become one of the most influential measurement papers of the 20th century. Stevens' paper, published in *Science* in 1946, was entitled "On the Theory of Scales of Measurement." The paper begins, as does this section of the chapter, with the story of the Ferguson Committee. The entire first page is focused on making connection with the issues that were of concern to the committee, and to Norman Campbell, an influential member of the committee. The Stevens paper is not, however, a head-on rebuttal of the committee's rather narrow definition of what constitutes scientific measurement. That firm corrections work would remain to be accomplished by the *axiomatic* representational measurement theorists during the last half of the 20th century, beginning with Luce and Tukey (1964) and their demonstration that additive conjoint measurement (ACM) could produce additive scales without concatenation.

Stevens' approach in 1946 was to shift the object of inquiry from concatenation operations to the identification and explication of various types of measurement scales. He begins, in the third paragraph of the paper, by agreeing with Campbell's relational theory of measurement, and then proposes ways to broaden it.

> Paraphrasing N. R. Campbell (Final Report, p. 340), we may say that measurement, in the broadest sense, is defined as the assignment of numerals to objects or events according to rules. The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement. The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale.

Clearly, Stevens' strategy here is not to directly oppose Campbell's definition of measurement as requiring concatenation, but simply to propose that it is only one of several ways to assign numerals according to rules. This is then followed up with twin questions of what other rule systems there might be, and what mathematical properties each would have. Stevens then proposes the practical aim of determining from the mathematical properties which statistical or mathematical analyses are logically admissible for each type of scale. In other words, his aim seems to be to present a few types of measurement scales, review their properties, and offer guidance regarding logically appropriate types of analysis for each. This pragmatic turn may be in large part responsible for the popularity and broad influence of Stevens' approach to RTM *scales*.

1.2.2.3 Stevens' Four Measurement Scales

The diagram in Figure 1.1 is a summary of Stevens' four measurement scales—nominal, ordinal, interval, and ratio—and their relationship to the "basic empirical operations" that characterize each. The simplest and least informative scale is the nominal scale on the left. It utilizes only one of the properties of numbers, identity (= and $\neq$), that is, whether two objects belong to the same category or not. This scale is essentially equivalent to what we refer to as "categorical" data. Stevens indicates that the nominal scale "represents the most unrestricted assignment of numerals. The numerals are used only as labels or type numbers, and words or letters would serve as well."

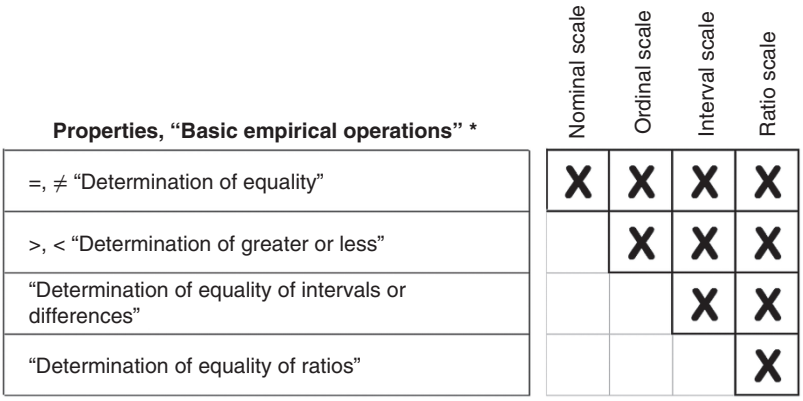


Figure 1.1 A summary of Stevens' four measurement scales.

Examples would be the numbers on jerseys in various sports such as hockey, baseball, and football, which are numerical. Non-numerical nominal scales could include professional or occupational groups such as medical doctors, attorneys, accountants, and so forth; or national groups such as Norwegians or Chinese.

Some non-numerical categories, such as college freshmen, sophomores, juniors, and seniors, are ordered categories. Senior standing is greater than junior standing, which is greater than sophomore standing, which is greater than freshman standing. These ordered categories would constitute an ordinal scale as shown in the second column of Figure 1.1, which adds a second property "determination of greater than or less than" (> and <). This is sometimes called the property of magnitude. In general, as we move from left to right (nominal to ratio) in the diagram the properties are cumulative, so an ordinal scale has the "equal/unequal" property of the nominal scale but adds to it the property of "greater or less than."

Other examples of non-numerical categories that constitute ordinal scales are Olympic medals where gold > silver > bronze or placement in a race where first place > second place > third place. Obviously, this last example of placement in a race could also be done with numbers, from 1 to whatever number is the last place, but also obviously the properties of the scale would not include additivity, since the number is only an ordinal ranking. This brings up the question of what kinds of mathematical operations make sense with ordinal scales. Stevens comments that:

> The *ordinal scale* arises from the operation of rank ordering. Since any 'order-preserving' transformation will leave the scale form invariant, this scale has the structure of what may be called the order-preserving group. A classic example of an ordinal scale is the scale of the hardness of minerals. Other instances are found among scales of intelligence, personality traits, grade or quality of leather, etc.
>
> As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety, the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply knowledge of something more than the relative rank order of data. On the other hand, for this 'illegal' statisticizing, there can be invoked a kind of pragmatic sanction: In numerous instances, it leads to fruitful results. While the outlawing of this procedure would probably serve no

> good purpose, it is proper to point out that means and standard deviations computed on an ordinal scale are in error to the extent that the successive intervals on the scale are unequal in size. When only the rank order of data is known, we should proceed cautiously with our statistics, especially the conclusions we draw from them.
>
> (Stevens, 1946, p. 679)

The major dividing line among these scale types is right in the middle of the four, between the ordinal scale and the interval scale. As Stevens says, it is at the interval level that "we come to a form that is 'quantitative' in the ordinary sense of the word. Almost all the usual statistical measures are applicable here unless they are the kinds that imply a knowledge of a 'true' zero point" (p. 679). Stevens uses the example of Fahrenheit and Celsius temperatures to illustrate the interval scale. Indeed, nearly all textbooks use this same example. That is because almost all other scales that possess interval properties also possess ratio properties. Lengths of tables are an example of a ratio scale. A table 4 m long is twice as long as one 2 m long. An interval scale would not have this property.

Temperature is a good example of an interval scale because it is so obvious that ratio comparisons cannot be made. Suppose that the high temperature in your town for yesterday was 35° and the high for today is 70°. You might be tempted to say that it is twice as warm today as yesterday, but when you convert the temperatures to Celsius, it now becomes 1.66° yesterday and 21.11° today, which would make it 12.7 times as warm today as yesterday!

In an interval scale, however, the intervals themselves, that is differences between magnitude values, *do* have meaningful ratio comparisons. For example, suppose there are two cups of water set at 78 °F. The temperature of the first cup of water is raised to 82 °F, and the temperature of the second one is raised to 80 °F. It could be said that the temperature increase for the first cup of water was twice as much as the second. The two intervals of change are in the ratio of two to one.

Narens (2002) gives a summary and critique (pp. 46–50) of Stevens' paper and appropriately entitles it "Stevens' Theory of Scales and Meaningful Statistics." He states that "Stevens, in his paper and other writings, is quite vague about what in general are the proper rules for assigning numbers to empirical objects—that is, what should be proper methods of measurement. He gives many specific examples, but this clearly is inadequate for the rigorous founding of a theory of measurement." This rigorous founding for a theory of measurement would come in another 15 years with the publication of the first volume of the *Foundations of Measurement* series (Krantz et al., 1971). Nevertheless, Stevens' place as a pioneer for broadening the approach to RTM in the human sciences is secure, clearly setting out for a wide audience how empirical relationships can be represented within, and isomorphic to, numerical structure.[8]

Scales are possible in the first place only because there is a certain isomorphism between what we can do with the aspects of objects and the properties of the numerical series. In dealing with the aspects of objects we invoke empirical operations for determining equality (classifying), for rank ordering, and for determining when differences and when ratios between the aspects of objects are equal. The conventional series of numerals yields analogous operations. We can identify the members of a numeral series and classify them. We know their order as given by convention. We can determine equal differences, as $8 - 6 = 4 - 2$, and equal ratios, as $8/4 = 6/3$.

8 Note the similarity of this paragraph from page 677 of Stevens (1946) with Michell's (1990) characterization of representational theories of measurement in his book *Introduction to the Logic of Psychological Measurement*: "According to the representational theory of measurement, the role of numbers in measurement is to represent. Numbers are assigned to empirical entities in such a way that certain relations between the numbers assigned represent empirical relations between the entities" (p. 29).

The isomorphism between these properties of the numeral series and certain empirical operations which we perform with objects permits the use of the series as a *model* to represent aspects of the empirical world (Stevens, 1946, p. 677).

The Stevens paper has exerted a major influence on the contemporary landscape of psychology and other social and behavioral sciences generally and has been incorporated into much contemporary statistical practice. Not only is his theory of scale types regarded as definitive by many within the behavioral sciences but it also stimulated interest in the study of scale types which eventuated in the development of the rigorous discipline of the axiomatic approach to *Relational Theory of Measurement* (RTM) from the 1960s until now. It could be said that Stevens' work with measurement scales represents an important contribution to the overall direction of the RTM, but not to the more thorough and conclusive *axiomatic* approach that is represented in the three volumes of *Foundational Measurement Theory* (Krantz et al., 1971; Suppes, Krantz, Luce, & Tversky, 1989; Luce, Krantz, Suppes, & Tversky, 1990).

1.2.2.4 Foundations of Measurement, the "Received View" of RTM

Narens (2002) gives a terse and cogent summary of the history of how the axiomatic approach to the relational theory of measurement and the three volumes of *Foundations of Measurement Theory* came to be, as illustrated in Figure 1.2.

> Measurement in its broadest sense consists of assigning numerically based entities to qualitative objects... This approach dates back at least to the foundational work on physical measurement by Helmholtz (1887) and Hölder (1901). It was also used by Hilbert (1899) to assign numerical coordinates to points of a qualitative Euclidean geometry. Much later, abstract versions of it for general measurement theory were given by Scott and Suppes (1958), Suppes and Zinnes (1963), and Pfanzagl (1959, 1968). The version of Scott and Suppes was later adopted as the theoretical basis for measurement in the seminal work on the subject, *Foundations of Measurement, Vols. I, II, and III* (Krantz, et al. 1971; Suppes, et al. 1989; and Luce, et al. 1990). Over the last three decades, this abstract version has dominated the theoretical measurement literature. For lack of better terms, I call this approach to measurement the *received representational theory*... (p. 211).

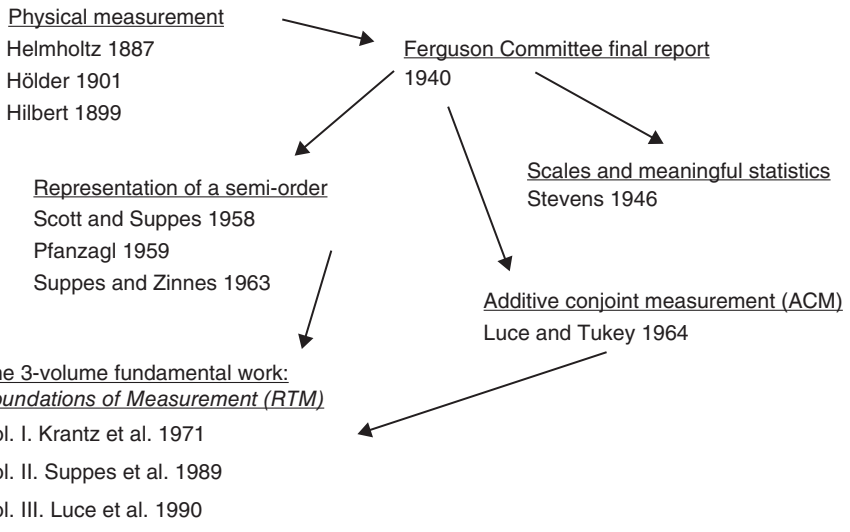


Figure 1.2 The development of RTM, the representational theory of measurement.

The early 19th century beginnings of measurement theory as the assigning of numerals to objects is clarified by Campbell's (1920, 1928) example of the isomorphism between empirical concatenation and numerical addition. Stevens takes a next step by defining the concept of measurement as "the assignment of numerals to objects *by rules*." He then makes an important strategic move by broadening the application of the concept with his statement that "the fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement." These three initial steps are followed by the axiomatic approach represented in the work of Scott and Suppes (1958), the work of Luce and Tukey (1964), and culminating in what Narens refers to as the dominant "received" view contained in the three volumes of *Foundations of Measurement* (Krantz et al., 1971; Suppes et al., 1989; Luce et al., 1990).

Quite abstractly, and with logical proofs and mathematics not accessible to many researchers in the behavioral or human sciences, this approach identified new classes of numerical structures that form the basis for measurement scales including more than the four types of scales popularized by Stevens. They then proved theorems of the existence of representations that mapped the numbers into one of the numerical structures thought to be appropriate to the empirical domain. They went farther and proved theorems of uniqueness that identified equally good parameters for representations into the mathematical structure, those that could handle data transformations as well. This is important. We don't want to hold our collective breaths to make sure we can use more than ordinal mathematics. We want at least additive structures so we can use least squares, correlations, factor analysis, and the like.

Patrick Suppes, the first author of the 1989 volume of Foundations of Measurement Theory, Volume II, earlier hoped this work would advance the field of practical calculations. Scott and Suppes (1958) remarked: "A primary aim of measurement is to provide a means of convenient computation. Practical control or prediction of empirical phenomena requires that unified, widely applicable methods of analyzing the important relationships between the phenomena be developed. Imbedding the discovered relations in various numerical relational systems is the most important such unifying method that has yet been found" (pp. 116–117).

Consider this spectrum that summarizes Scott and Suppes' two requirements for practical control, first, the use of widely applicable methods of analyzing the important relationships between the phenomena, and then that the discovered relationships be imbedded in various numerical relational systems. These are the two sides that need to be brought together, the empirical on the left and the numerical on the right. We have seen with Campbell's simple beam balance example how the empirical-to-numerical connection can be illustrated and made possible by concatenation:

Constructs and "would-be" scales from qualitative reality ↔ Imbedding (injecting)[9] the relationships found into a numerical system

A student of Suppes, Zoltan Domotor, wrote a chapter in the book *Philosophical and Foundational Issues in Measurement Theory* (Savage & Ehrlich, 1992). It cut deeply into the critique of axiomatic RTM and was given the position of the final chapter in that book. In it, he advances the position that of the two sides of the spectrum drawn above, the imbedding side cannot supply all of the needed theory. He called the left-side projection from theory. He argued that projection must take precedence and must precede imbedding, "...my conclusion will be that representation or injection of the qualitative into a quantitative is fully justified only when coupled with the dual

9 Scott and Suppes call this mathematical expression "imbedding" while Domotor calls it "injection."

procedure of projection" (p. 196, italics ours). By projection he means "The idea is that theoretical models are projected onto their observational structures..." (p. 195).

Domotor is framing an argument as a measurement philosopher. He does not give technical procedures for how to project theory onto an observational structure nor how to get either the theory or the models that tell us what to observe. Scott and Suppes in the quote above also did not give much attention to how we reach into the qualitative reality and extract theoretical constructs and turn them into scales. They talked about the use of "widely applicable methods of analyzing the important relationships" but then zeroed in on imbedding them into a numerical structure. Right from the beginning of his article, Domotor seems as committed to quantitative scales as Scott and Suppes but reminds us that both sides are needed.

Domotor gives his readers, including those who might be psychometricians, a strong statement of what needs to be done first. Psychometricians with broad experience developing instruments have learned how to identify and define empirical "constructs," find observable indicators of them, and thus create scales. As psychometricians, we could name the left side as "design and develop constructs and scales from qualitative reality." In Chapters 4 and 5, we will explain how analyzing empirical domains and developing domain theory can lead to models that can be used to "project," that is to design and develop instruments that tap into observational structures (Mislevy, Almond, & Lucas, 2004; Mislevy & Risconscente, 2005).

Psychological instruments are analogous to the extravagant instruments used at the Olympics but more modest in scope. It is useful to note that both types of instruments can provide the means for enabling users to "make contact" with tasks calibrated on scales where the instrument delivers a number that can be operated on by psychometric procedures. Also, the order of the tasks, or the order of physical or mental attainments in the theory, has a meaning that can be used in that all-important aspect of validation, "interpretation." That contact, together with enough such tasks for reliability, makes possible the issuing of a score after statistical work. The meaning comes from the theory and from the validating evidence that there is some reality in what people are doing with their minds or in their physical performances—an invisible mental or physical reality that can be captured by scores from an instrument.

Domotor (1992) objects to the empiricist approach and promotes a realist position. The theory is more than a language blessed with abstract terms to bandy about, nor is it just to make predictions. There is no reason to believe that the world contains anything that corresponds to the abstract terms of empiricism. Instead, he argues, "...a theory is more than just an instrument for driving predictions; it must satisfy our demand for an explanation by telling us how and why things happen" (p. 196). However, for this purpose, the theory must aim at a genuine description of what is happening. The practice of science shows us that it is precisely by viewing theories realistically that science progresses. It is by thinking of theories as literal albeit perhaps approximate descriptions of what goes on that we can suggest good ways of extending or generalizing them and can see where they might require modification.

The "received view" of RTM in the Krantz et al. three volumes has not been without comments, suggestions, and critique, as exemplified in Domotor's call for a realist approach to understanding measurement. Others have pointed out and objected to the proliferation of axioms, proofs, and theorems in the Krantz et al. version of RTM that in and of themselves lack practical utility. Domotor's pithy comment on this, after giving 12 pages explaining logical reasons for the "troubling profusion of representation results" was, "I suggest to project first and represent later" (p. 214).[10]

10 This is an intriguing thought that perhaps more clarification on the "qualitative reality" side could simplify the task of proving numerical representation and uniqueness.

However, what such theorems do is provide foundational assurance that our number system assumptions are justified. It would be foolish to not recognize the achievement of the pioneers who have given the world Foundations of Measurement Theory, but this work was not designed to replace the tried-and-true methods of true score test theory. That possibility has been suggested by some (Michell, 1990, 2020), but as Cliff (1992) observed 30 years ago the expected axiomatic RTM revolution in psychometric theory and methods had not occurred by 1992, and 30 years later it still hasn't. The discipline of psychometrics continues to move forward with practical and useful test theory methods for the most part unaffected by RTM. We should be patient, however, with those who continue their difficult RTM work in hopes of possible useful convergences.

Fortunately, the axiomatic RTM theorists are self-critical. Luce and Narens (1994) published the results of research funded by a National Science Foundation Grant. In it, they listed ongoing work, various unsolved problems with RTM, and several suggested new directions. They listed 15 problems requiring additional work. Among them are two of the most important issues for our key interests in this book: the practical work of developing actual measurement scales and clarifying the meaningfulness issue.

One useful and practical result to come from mathematical RTM is Luce and Tukey's (1964) demonstration of the power of ACM. When its axioms are met, it produces an interval scale without the operation of concatenation. If an instrument can be shown to meet the axioms of ACM, and the mathematical model fits the data well enough to assume conditional independence, investigators have evidence that they are justified to treat their constructed scale as additive with a convenient zero and a unit that is as persuasive as the unit and zero created for the Fahrenheit and Celsius temperature scales.

1.2.2.5 The Contrast Between Foundations of Measurement and the Science of Applied Psychometrics

At the end of Section §5, we indicated that we would only touch lightly on the "logical, theoretical, and philosophical foundations of measurement." We took a brief look at the history and principles of representation and meaningfulness as embodied in the RTM and more specifically, the axiomatic approach to the RTM as it emerged in the mathematical psychology literature of the past 50 or 60 years.

This is not to be confused with CTT and its derivatives, which began much earlier—at the beginning of the 20th century—which constitute the foundational practical science upon which the past century of psychological testing is based. With our brief look at RTM now completed, we turn to this dominant approach to psychometrics throughout the remainder of the book—so-called "true score theory," which includes CTT, item-response theory, and the myriad methods that have grown out of them. The axiomatic approach to RMT has had very little impact on psychological testing or psychometric practice. It consists primarily of logical clarification, conceptual demonstrations, and axiomatic logical proofs which arose in mid-century as an effective rebuttal to the accusation from some quarters of the scientific community that measurement in psychology, and indeed in all the behavioral sciences, is at best unscientific and at worst perhaps meaningless.

Each of these two traditions is referred to as "measurement theory" as shown in Figure 1.3 but they differ substantially in their approach, in their historical roots, and in their practical impact. As shown in that figure, both kinds of measurement theory have both a theory aspect and also a scaling aspect. The measurement theory aspect is foundational and formal and deals with the mathematical and conceptual/theoretical conditions under which scales can be constructed, while the actual process, the practice by which numbers are assigned to objects, or attributes of objects, is called scaling.

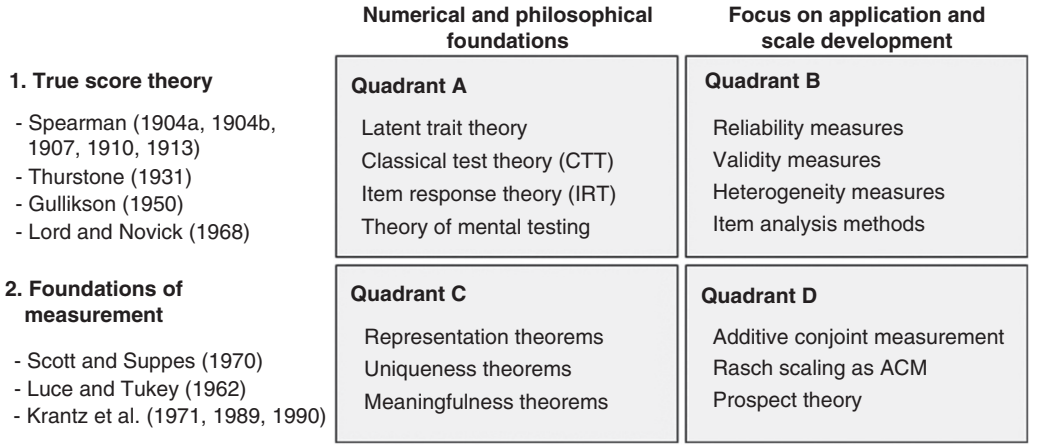


Figure 1.3 A comparison of two contrasting approaches to psychological measurement: (1) True Score Theory, the traditional psychometric testing tradition, with its myriad derivative developments; and (2) the Foundations of Measurement axiomatic approach to the RTM, the current "received view" of RTM.

True score theory, Quadrant A in the table, is essentially the main line of theory and practice in psychometrics as it has grown over the past 12 decades. It includes the majority of what is explained in the remainder of this book. It begins with the emergence of factor analysis at the beginning of the 20th century and the subsequent developments of the theory and mathematics of measures of reliability, validity, test homogeneity, and a wide variety of other psychometric equations and processes—fueled by Pearson's (1896) correlation coefficient, his discovery of the mathematical basis of factoring (2001), and Spearman's introduction of factor analysis. Most of these date back to the first half of the 20th century. These historical developments are described well by Gulliksen (1950), with many references to early fundamental sources such as Spearman (1904a, 1904b, 1907, 1910, 1913), Thurstone (1931), and many others. Gulliksen acknowledges that his book draws heavily upon the teachings of Thurstone "and particularly upon his now unavailable book *Reliability and Validity of Tests* (Thurstone, 1931)."

Latent trait theory is also a part of Quadrant A and includes the early sources on factor analysis such as those by Thurstone (1935, 1947), Harman (1960), and Horst (1965), that chronicle the development and elaboration of a wide variety of factoring models. These early psychometric efforts are referred to collectively as CTT (classical test theory) to distinguish them from modern test theory, that is, item response theory (IRT). CTT as described in exemplary texts such as Lord, Novick, and Birnbaum (1968) and Allen and Wen (1979) was first codified by Novick (1966). The definitive source explicating and summarizing these quantitative foundations of psychometrics over the first 70 years of its existence is the 1968 Lord and Novick text *Statistical Theories of Mental Test Scores, with Contributions by A. Birnbaum*. Probably no other single book since that time has had a comparable impact upon the past century of psychometric practice. It is a one-volume sourcebook on the statistical/mathematical basis of CTT supplemented by an introduction to the beginnings of item-response theory in Chapters 17–20, written by Allan Birnbaum.

Quadrant B consists of the many psychometrically useful analytical practices and tools that grew out of the numerical and philosophical foundations of CTT and IRT, such as those presented in the Gulliksen book or the Lord and Novick book, including Jöreskog's (1969) epoch-making maximum likelihood method of confirmatory factor analysis (to be covered in Chapter 6 and Computational

Guide 1). It also consists of the multitude of specific psychological tests and measures that have been produced using these practices and tools. This would include the over 14,000 specific testing instruments chronicled in the 21 published editions of the *Buros Mental Measurement Yearbooks* from their inception in 1938 down to the most recent edition in 2021 (Carlson, Geisinger, & Jonson, 2021). This was the golden age of test development, still in progress.

Quadrant C is the axiomatic approach to the RTM. It began in the 19th and early 20th century, as was presented in Figure 1.2 and the accompanying pages, and culminated in the three volumes of *Foundations of Measurement,* the currently accepted method of RTM. This approach had its beginnings in the mathematical psychology literature of the 1960s and 1970s (Coombs, Dawes, & Tversky, 1970), as a natural follow-up to the Ferguson committee controversy, Stevens' gentle rebuttal in his 1946 measurement scales paper in *Science*, and the surprising and convincing demonstration by Luce and Tukey (1964) of rigorous psychological measurement without concatenation using simultaneous conjoint measurement. Stevens' four measurement scales have become second nature to the behavioral science disciplines and could be considered to also be a variety of RMT (the Relational Theory of Measurement), but not of the *axiomatic* approach to the Relational Theory of measurement that now characterizes the mathematical psychology literature. Clearly, Stevens' four measurement scales do not constitute a rationally defensible justification of the fundamental principles of measurement but they have been a useful part of psychometric progress.

In addition to ACM, the poster child of axiomatic RTM, there have been many Quadrant D mathematical models loosely connected to axiomatic RTM developed over the past century, including signal detection theory (SDT), other psychophysical models, information theory, unfolding theory, mathematical learning theory, scalogram analysis, multidimensional scaling, Thurstone's law of comparative judgement and his law of categorical judgement, models of preferential choice, utility theory, etc. (Coombs, 1964; Torgerson, 1966). Perhaps one of the most notable achievements to grow out of axiomatic RTM was the behavioral economics work of Tversky and Kahneman. Daniel Kahneman was awarded the Nobel Prize in Economics in 2002 for his part in the development of prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1974), just six years after the death of his research partner Amos Tversky, in 1996. A brief axiomatic treatment of prospect theory was given in the appendix of Kahneman and Tversky's (1979) paper, with acknowledged assistance from David Krantz, the lead creator of the three foundations of measurement volumes. Perhaps there is also a tie-in between axiomatic RTM and Rasch (1960) scaling, a precursor of item response theory. There is some evidence that the Rasch scale fulfills the conditions of ACM (Perline, Wright, & Wainer, 1979; Karabatsos, 2001; Newby, Conner, Grant, & Bunderson, 2010).

1.2.2.6 Metrology and Measurement Units

While the Ferguson Committee was engaged in its seven-year debate, failing to consider forms of measurement other than those used fruitfully by physicists, engineers, and chemists, International Metrology[11] was becoming more and more international, and developing cooperation on setting standards and promulgating a System of Units (SI), a metric system that scientists of all stripes could use worldwide. It led to great economic and social gains, through the application of science-based standards throughout industry and the professions. Metrology is the science of

11 International metrology has been with us for over 200 years. The early metrologists received a major boost for their "meridian-arc" definition of the meter as one ten-millionth of the earth's circumference following the completion of Méchain and Delambre's six-year survey of the distance from the equator to the North Pole in 1798 (Ferreiro, 2011; Alder, 2021)

measurement. As philosophy could be considered "thinking about thinking," metrology could be considered "measuring measurement" (Brown, 2021). It is much more mathematical than other approaches to measurement. It has three divisions, Scientific Metrology, Applied (or industrial) Metrology, and Legal Metrology. It has made remarkable advances up through 2019, during which time all physical artifacts for measuring length, and finally in 2019, calibrating weights, were replaced. Physical artifacts are not precise enough—not easily calibrated nor traced. International Metrology is not usually explained in textbooks of this sort, but it is so important to the future of any scientific field that we have prepared a Chapter 8 entitled "Metrology: Scientific, Applied, and Legal.

In this book's introduction, we promised to show that while a demonstration of concatenation is *sufficient* to establish a measurement scale, it can be shown to not be a *necessary* condition for scientific measurement. This is most clearly shown in the stunning achievements of contemporary metrology scientists. In May 2019, the International System of Units (SI) replaced the last physical object suitable for concatenation demonstrations. It was THE KILOGRAM (called the IPK [International Prototype Kilogram]) that was carefully stored under ultra-safe and controlled conditions in a metrology laboratory in Paris. It has been retired and replaced by an average of new international estimates of fundamental constants, following the latest experimental estimates of the Planck constant, averaged from nine Metrology Laboratories in several countries to a relative uncertainty of 1.2×10^8 decimal places.

This is a profound achievement in the broad science of measurement, retaining the familiar physical measures such as meters and grams, but tying them to seven fundamental constants of the physical universe. And it comes with a refreshing openness and encouragement for new approaches to measurement in every conceivable domain. Who can imagine what its effects might be on the fledgeling science of psychometrics in the coming decades?

You can read more about this amazing attainment in worldwide cooperation in Stock, M., Davis R., de Mirandes, E., and Milton, M. J. T. (2019), "The revision of the SI—the result of three decades of progress in metrology." The abstract of this *Metrologia* paper reads as follows:

> On 16 November 2018, a revision of the SI was agreed upon by the General Conference on Weights and Measures. The definitions of the base units were presented in a new format that highlighted the link between each unit and a defined value of an associated constant. The physical concepts underlying the definitions of the kilogram, the ampere, the kelvin and the mole have been changed. The new definition of the kilogram is of particular importance because it eliminated the last definition referring to an artefact. In this way, the new definitions use the rules of nature to create the rules of measurement and tie measurements at the atomic and quantum scales to those at the macroscopic level. The new definitions do not prescribe particular realization methods and hence will allow the development of new and more accurate measurement techniques.

Figures 1.4 and 1.5 are photocopies of tables 1 and 2 taken from page 128 of BIPM (2019), the Bureau International des Poids et Mesures (English text version in Newell & Tiesinga, 2019, *The International System of Units (SI), ninth edition*). Figure 1.4 shows the seven standard constants, together with the symbols, numerical values of the constants, and the defining unit of each. Figure 1.5 gives the base quantity of each of the seven familiar base quantities derived from each constant, the typical symbol, and the base unit name of the measuring units for each (second, meter, kilogram, etc.) with its symbol.

Defining constant	Symbol	Numerical value	Unit
Hyperfine transition frequency of Cs	$\Delta\nu_{Cs}$	9 192 631 770	Hz
Speed of light in vacuum	c	299 792 458	m s^{-1}
Planck constant	h	$6.626\,070\,15 \times 10^{-34}$	J s
Elementary charge	e	$1.602\,176\,634 \times 10^{-19}$	C
Boltzmann constant	k	$1.380\,649 \times 10^{-23}$	J K^{-1}
Avogadro constant	N_A	$6.022\,140\,76 \times 10^{23}$	mol^{-1}
Luminous efficacy	K_{cd}	683	lm W^{-1}

Figure 1.4 The seven defining constants of the *SI* and the seven corresponding units they define, a photocopy of table 1 from page 128 of *The International System of Units (SI)*. This SI Brochure is distributed under the terms of the Creative Commons Attribution 3.0 IGO (https://creativecommons.org/licenses/by/3.0/igo/).

Base quantity		Base unit	
Name	Typical symbol	Name	Symbol
Time	t	Second	s
Length	l, x, r, etc.	Metre	m
Mass	m	Kilogram	kg
Electric current	I, i	Ampere	A
Thermodynamic temperature	T	Kelvin	K
Amount of substance	n	Mole	mol
Luminous intensity	I_v	Candela	cd

Figure 1.5 The seven *SI* base units and their names and symbols, a photocopy of table 2 from page 128 of *The International System of Units (SI)*. This SI Brochure is distributed under the terms of the Creative Commons Attribution 3.0 IGO (https://creativecommons.org/licenses/by/3.0/igo).

1.2.2.7 Measurability in the Behavioral and Social Sciences

Psychologists often speak of nomological nets that provide linkages among a variety of theoretical constructs. This attainment in the realm of physics and the multiplicity of types of experiments in the various metrology labs give a deep example of how advanced nomological nets can become, and how long it can take, even in physics. Concatenation can still be performed because each country can create its own standard prototype kilogram—or any other unit. But it cannot be said that a demonstration of concatenation is necessary to prove the measurability of any standard unit in the SI. The references above also provide a very clear common-sense understanding that physical phenomena that have no volition and answer only to unchanging natural laws can be measured with such incredible precision that it would be ludicrous to require that standard of accuracy to phenomena in the human-created domains wherein behavioral measurement takes place.

Often only order is necessary to have explainable theories that are consistent and useful, and are backed by solid constructs, good instruments, and good implementation techniques to reduce construct-extraneous sources of error. But as shown in the axiomatic RTM sub-section earlier, with ACM, when people's scores and a set of tasks are conjoined, we can establish additivity as well. But, ratio scales are not always to be expected nor needed in human-made domains that possess the numerical values appropriate for instrument development.

1.2.3 (§7) *Scientific* and *Stochastic* Variables: When May "Data" Be Treated By Statistical and Test Theory Methods?

The genius of CTT is to conceptualize test scores as consisting of two components: the "true score" which can be thought of as the deterministic component, and "measurement error," which can be thought of as the stochastic/probabilistic component—a random variable. Random does not mean haphazard—far from it. A random sample, for example, is rigorously representative of the population from which it is drawn. That is, each member of the population has an equal chance of being included in the sample. It is an exacting process to create a "sampling frame," a list that includes all in the population, and to then use a truly random process, something like a random numbers table, to select the sample.

Stochastic variables can provide an important kind of leverage. A random variable, although unpredictable in a single observation, is predictable in the aggregate. Suppose, for example, you wish to stochastically model the distribution of the outcomes, the sums of values, in tossing two dice. In other words, you wish to create a *probability model* of outcome sums. Obviously, one could not predict any particular outcome of tossing two dice, but if the dice are truly unbiased and random, one could specify with precision and exactness what the distribution of outcomes would be across many such tosses, as shown in Figure 1.6 which constitutes a kind of "diagrammatical demonstration." Each of the 36 possible rolls of the 6 possible outcomes on one die and the six on the other have an equal probability of occurring. Each of these possible, and equally likely, two-dice rolls for each sum value are grouped above the appropriate sum value, creating a systematic distribution. Each of these 11 sum values (from "2" to "12" inclusive) has a precise probability as shown in the probability vector above the stacks of outcomes. These could be considered to be population outcomes, but they are more than that. They are ontological probabilities that capture the true and timeless nature of the mathematics of dice tossing for samples of size two.

What has been created here is a *sampling distribution of sums*. It is essentially equivalent to a sampling distribution of means (SDM), except that we have not divided the sums of the two-dice tosses by *N*. If we now had the time, the inclination, and the patience, we could create similar sampling distributions for three-dice tosses ($N = 3$), which would have sixteen unique sum values along the *x*-axis, the abscissa (with sum values ranging from 3 to 18), and 216 three-dice outcome permutations; or even for four-dice tosses ($N - 4$), which would have an abscissa with 21 unique sum values (4–24), and 1296 four-dice outcome permutations. And, even though this sampling distribution with $N = 2$ is triangular in shape, with increasing sample size (N) from 3 and up, the sampling distribution of sums will become more and more normally distributed, that is, more Gaussian. This is in accordance with the *Central Limit Theorem,* which holds that (regardless of the shape of the

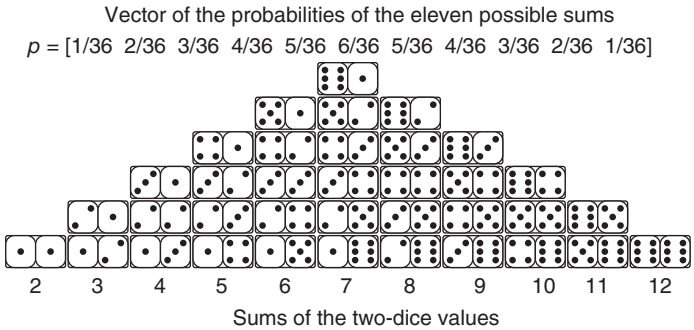


Figure 1.6 A graph of the sampling distribution of the 36 possible paired-dice tosses and their respective probabilities.

population distribution) as sample size increases the sampling distribution of means, the SDM, will converge to Gaussian. The same is true of the sampling distributions of sums and of proportions.[12] In other words, a bell-shaped Gaussian distribution is the limiting form. In Appendix 2, methods are introduced for calculating descriptive statistics for probability models. Means, variance, and other statistics can be used to describe with statistical precision the nature of distributions, spatial patterns of probability densities, and so forth. CTT consists of a variety of ways of managing and reducing error and discovering "true score" structure when an error is present. When applied to measurement error, rather than dice tossing, probability models enable us to characterize, estimate and reduce measurement uncertainty with precision as demonstrated in Section §13. Hence the value of stochastic modeling in psychometrics.

A theoretical probability model, like the one just described, has three necessary conditions: first, the outcomes in the model are *mutually exclusive*, second, the outcomes are *exhaustive*, and third, the *probabilities sum to 1.00*. In Figure 1.6, the outcomes are clearly mutually exclusive: if the sum of two dice is 7, it cannot also be 8 or 9. The outcome possibilities are also exhaustive: there are no other possible sums of two dice than the eleven sums specified. The reader can also verify that the eleven probabilities in the probability vector above the diagram do indeed sum to one. This sum of 1.00 is another way of stating the property of *necessity*—each one of the 11 sum-value outcomes and their associated probabilities are necessary to the complete model. Given the verification of these three properties, we can conclude that this sampling distribution of sums is indeed a probability model.

The important part that random variables play in psychometric theory and CTT will not be fully explained here but will become clearer as we progress throughout the remaining chapters. This is enough to establish that there is indeed explanatory power in establishing a variable as a stochastic variable, a true random variable.

1.3 Statistical Methods: Location, Dispersion, and Statistical Inference

1.3.1 (§8) Group Level of an Attribute—The Mean Location of the Group's Scores on an Interval Scale

We will now demonstrate the calculation of indices of central location and dispersion of scores using fictional quasi-Galtonian data as if measured in a sample of N pairs of fathers and their firstborn sons. If we accept the procedure of measuring head circumference to be a scientific variable recognized as a stochastic variable x, then its representation conditions enable measurement on an interval-metric scale. Hopefully, it can be seen then that the arithmetic *mean* is a very informative characteristic, at least intuitively, if one wants to compare the overall levels of two groups of people in two data sets of a purely quantitative nature. Besides the arithmetic mean, there are other useful characteristics that express location—the median and the mode. All three are expressive of the overall tendency of a data set to be "located," or its position to be "centered," somewhere on a scale. So, quite naturally they are referred to in statistical terminology as *characteristics of location* or *measures of central tendency*.

Consider a data set of fictional head circumference data for a sample of $N = 30$ fathers and also for their sons (one son each, so also $N = 30$). Suppose that we also round the continuous-value head

12 This is not only true of means and sums, but also of proportions, since the proportion is a special case of the arithmetic mean. A proportion is the mean of a binary variable. Note, however, that the rolling of dice or the tossing of coins creates discrete distributions, while the Gaussian distribution by definition is continuous. However, although the sampling distribution created by coin tosses is technically a discrete binomial distribution, it is closely approximated by a Gaussian distribution as N becomes large. (See Brown, Hendrix, Hedges, and Smith (2021), pages 43 to 58, the section entitled "Classical probability theory and the binomial: The basis for statistical inference.")

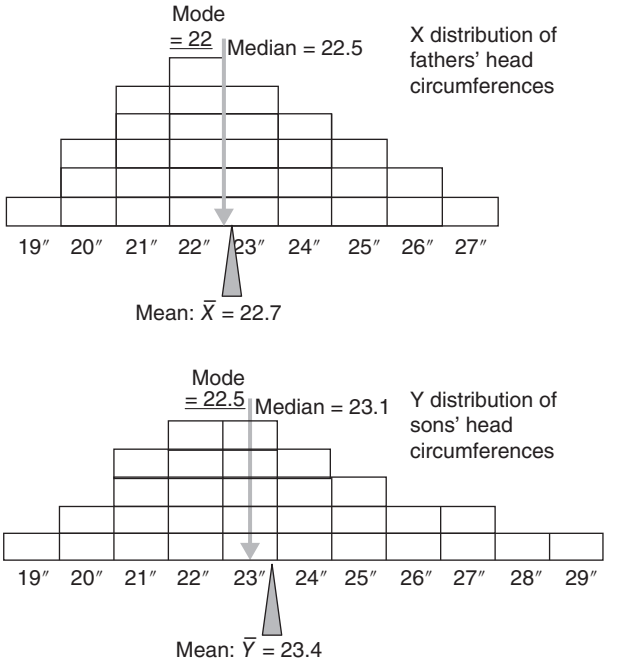


Figure 1.7 Histogram comparison of fictional head circumferences of 30 fathers with corresponding fictional head circumferences of their sons, with the mean, median, and mode shown for each distribution.

circumferences to the nearest inch. The values measured are pictured in inches as points on an interval scale in Figure 1.7. The example will be simplified and more illustrative working with rounded discrete data. If we represent the data for fathers as 30 wooden blocks, each block representing the head circumference value of one of the 30 fathers, and each stacked above its appropriate integer-rounded circumference value along a numerical scale from 19″ to 27″, as shown in the upper half of Figure 1.7, it creates a *histogram* revealing the shape of the distribution of head circumferences for fathers. The same is done for the head circumferences of sons, as shown in the lower half of Figure 1.7 Each value within each distribution has a "relative frequency," or probability. For example, the value 27″ in the fathers' distribution contains only 1 of the 30 blocks, so its relative frequency or "probability"[13] is 1/30. Similarly, the value 24″ has a probability of 4/30 within this sample, and the value 22″ therefore has a probability of 6/30, etc.

We can find the average value (also called the *mean*) of the 30 scores of the distribution for the fathers, by first summing all thirty of the observed values, which comes to 681, and then dividing the 681 by 30, to get a mean of 22.7.

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$\overline{X} = \frac{\begin{array}{l}(19 + 20 + 20 + 20 + 21 + 21 + 21 + 21 + 21 + 22 + 22 + 22 + 22 + 22 + 22 \\ +23 + 23 + 23 + 23 + 23 + 24 + 24 + 24 + 24 + 25 + 25 + 25 + 26 + 26 + 27)\end{array}}{30}$$

$$= \frac{681}{30} = 22.7$$

13 We are using the term *probability* descriptively here, rather than ontologically, to simplify the demonstration of the conceptual basis of the calculation of fundamental statistics.

We use the symbol "X-bar" $(\overline{X})$ to represent the mean of distribution X. The capital sigma sign (Σ) in the numerator of the formula has an "$i = 1$" at the bottom of the sigma and an "N" at the top of the sigma. These indicate that you are to sum all the X_i values, from the first (with subscript = 1) to the last (with subscript = N). This sum is then divided by N. A simpler notation, and a more efficient way to get the mean for this sample is to obtain the "sum of products" of each of the X_i values multiplied by its own probability p_i, as explained in Appendix 2.

$$\begin{aligned}\overline{X} &= \sum(X_i p_i) \\ &= 19\left(\frac{1}{30}\right) + 20\left(\frac{3}{30}\right) + 21\left(\frac{5}{30}\right) + 22\left(\frac{6}{30}\right) + 23\left(\frac{5}{30}\right) + 24\left(\frac{4}{30}\right) \\ &\quad + 25\left(\frac{3}{30}\right) + 26\left(\frac{2}{30}\right) + 27\left(\frac{1}{30}\right) \\ &= \frac{19}{30} + \frac{60}{30} + \frac{105}{30} + \frac{132}{30} + \frac{115}{30} + \frac{96}{30} + \frac{75}{30} + \frac{52}{30} + \frac{27}{30} = \frac{681}{30} = 22.7\end{aligned}$$

The mean of Y, the distribution of the head circumferences of the sons, can be similarly obtained.

$$\begin{aligned}\overline{Y} &= \sum(Y_i p_i) \\ &= \frac{19}{30} + \frac{40}{30} + \frac{84}{30} + \frac{110}{30} + \frac{115}{30} + \frac{96}{30} + \frac{75}{30} + \frac{52}{30} + \frac{54}{30} + \frac{28}{30} + \frac{29}{30} \\ &= \frac{702}{30} = 23.4\end{aligned}$$

Calculations such as this can also be done somewhat more elegantly and systematically using matrix multiplication, as shown in the first two paragraphs of Appendix 3.

The mean is an agglomerative characteristic of data distribution, with two interpretations. The first is empirical, as a joint level of an attribute (such as head circumference) that is common to a group of people. That is, it summarizes the central tendency location of the group. The second interpretation is formal, as a point representing a mathematically central location of a set of numbers on a scale. The mean can also be thought of as the "expected value," in the sense that it gives the value that is "expected," that is, most likely for an unknown score, as explained in Appendices 1 and 2. It can also be thought of as the balance point for the distribution of scores as can be observed for the two distributions shown in Figure 1.7, where a small "wedge" or fulcrum is shown at the location of the mean for each of the distributions. If we were to model one of these distributions as an actual physical object with 30 wooden blocks stacked on a ruler or scale of negligible weight, the mean would be the precise location where a wedge must be placed to perfectly balance the distribution. As such, the mean is obviously highly affected by extreme values. If, for example, you were to take the block at location "27" on the X distribution, the highest score, and move it out to a location of "36," it would shift the balance point, the mean, from 22.7 to 23.0.

A typical textbook list of major descriptive characteristics of the central tendency location of a data set includes, besides the mean:

the *median*, defined as the middle value of the scores when ordered by magnitude, that is, as the point along the scale that bisects the area of the curve; and
the *mode*, the value with the highest frequency, so, in a sense, its "typical" value.

The mode as a physical characteristic of the stack of blocks would be the location of the highest stack of blocks, which for the X distribution is 22, with a frequency of 6. For the Y distribution, there

are two scores, 22 and 23, that each has a frequency of 5, so the mode is in the center of them, with a value of 22.5. See Appendix 1 for an account of the properties of mean, median, and mode, and their calculation.

1.3.2 (§9) Comparing Median vs. Mean: Which One Is "Fair"?

Let's go back and compare the mean $\overline{X}$ vs. the median with respect to their differing behavior when the data undergo changes or move along the scale. The mean, as the sum of the individual values divided by N, takes the magnitude of every score into account, whereas the median does not. The mean reacts to any small change even in one value—as we can recognize by recalling the formula again:

$$\overline{X} = \frac{(x1 + x2 + \cdots + xi + \cdots + xN)}{N}$$

Therefore, the mean as a measure of central tendency operates as a center of gravity of the data points and their distribution along the scale. So, the mean could be "computed" mechanically by balancing a measuring stick to equilibrium when the negative deviation of the data points to the left are counterbalanced by positive deviations to the right. Here we have an analogy with the first mechanical moment—balancing the data set on the scale to equilibrium means that the scale, as a "lever," is supported at the data center so that the sum of the deviations in magnitude from left to right, the first-order central moment, is zero.

Perhaps the greatest advantage of means as a measure of central tendency is that they fit a Gaussian distribution as the sample size N over which they are summated becomes large. This is the *Central Limit Theorem* discussed in Section §7. The *Central Limit Theorem* does not apply to medians.

The median is not as sensitive and does not respond to small moves of an individual data point until it jumps over the original median to the other side, i.e. only if the value changes its ordering relation with respect to the original median. Therefore, we could say that the median is not so "fair" as to represent the influence of each individual data value in relation to the overall central characteristic of the data set: Imagine how a change in one subject's individual attitude can influence the overall attitude of a group of people if the "composite group attitude" were measured by the arithmetic mean on the one hand or by the median on the other. The insensitivity of the median to individual data points is sometimes useful as it is not so influenced by just one or a few extreme values falling far outside the typical range. Measuring central tendency with the median rather than the mean could be a reasonable policy in some particular areas of research. For example, in reaction time studies there are occasionally extreme outliers when a distraction can make one reaction time many times larger than the rest. This could create misleading comparisons if means were used. The problem could be avoided by using medians. In statistical terminology, it is said that the median is robust with respect to outliers. It can also be useful to compare the mean and median of the same data set to learn what happens when a certain non-typical trial or nontypical subject (or a couple of subjects or trials) is removed.

1.3.3 (§10) A Psychometric Look at the Mean $\overline{X}$, Item "Difficulty," and p

Many behavioral and psychological attributes of people as subjects in a research study are observed through diagnostic methods that are psychological "tests" in the narrower meaning of the usual common-sense usage, such as admission tests, etc.—*not* in the broader sense, which calls *any*

diagnostic method a test in the special language of Test Theory. A lot of the "real" tests may be called "performance tests" in a general sense as we speak about academic performance in our rather competitive civilization. Such a test, whose scores are formalized as research variables and properly represented by corresponding random variables X, or Y, ..., typically stands for a real practical task to be solved by the respondent, and the *test score*, x or y, then expresses a quantitative level of "how well the test task has been performed." That is, it is an index of the subject's performance. Then, unlike the mean of somatic measurements in Galton's studies, a mean of the scores of such a test y may also be interpreted as an overall *level of performance* $\overline{Y}$ of a particular group of persons. Also, we can compare the means of two groups of people with this underlying performance interpretation. For example, we can compare "performance" in a simple clinical test of verbal ability, let's call it the *V-test*, in two groups of patients with two different medications for dementia. Let's denote the score on the *V-test* by the random variable Y—it consists of the task to produce as many words as the patient is able in one minute, namely the words which start with a given letter, or names of animals, and so on.

In the history of psychometrics, the oversimplified explanation of a *performance score* was based on two concepts: the *ability* of the tested subject and the *difficulty* of the task he or she was supposed to perform. If the task has only two alternatives, performed or not performed—it means accomplished or failed—the "equation" of such a performance contained just a single *dominance relation* between the two members:

$$\begin{aligned} &\text{Abilitylevel} > \text{task difficulty} \quad \Rightarrow \quad \text{Accomplished, that is, performance is "good"} \\ &\text{Abilitylevel} < \text{task difficulty} \quad \Rightarrow \quad \text{Failed, that is, performance is "bad."} \end{aligned}$$

But even quantitative scores of the individual's performance are related to the difficulty level of the task. Performance and difficulty are complementary and counterbalance each other. Therefore, finding that Group 1 has a higher mean score than Group 2, that is, $\overline{Y}_1 > \overline{Y}_2$, could be interpreted as the higher ability level of Group 1, but also as an indication of various conditions which might have decreased the level of test difficulty in the testing session for Group 1.

Of course, the rigorous conditions of standardization should not allow substantial influence on test difficulty, but the test could be more or less difficult for given categories of population specified by age, sex, etc. So, it is important to remember that an unspecified "mean level $\overline{Y}$" can be interpreted in two ways: namely either as an agglomerate *attribute of the group of people* as related to a given test or as an attribute or *property of the method—of the test* with respect to the given sample or population.

The usual clinical viewpoint of "mean level $\overline{Y}$" evaluates and compares particular respondents through their test score representations considering them as attributes of *human qualities* that need to be treated, and possibly improved. On the contrary, from the psychometric point of view, the same raw materials, namely the test scores, are used to study and compare various modifications of a test and its properties as *characteristics of a diagnostic procedure* that are to be improved to offer better practical and scientific information.

Thus, by having a standardized test administered on a representative and large sample or, still better, after replications on several such samples, we can learn about special psychometric indices as indicators of the overall diagnostic quality of the test. Such a *test calibrating study* can yield important information about the psychological test or test battery and suggest practical modifications for its diagnostic improvements.

Information about the *appropriate difficulty* of a test for a certain category of the population, i.e. age, gender, education, etc. is one of the basic properties of any test. In its simplest case, the mean

score, $\overline{Y}$ is used as a *psychometric index* of *test difficulty*. It is one of the important diagnostic properties of the test.

Later we will introduce several other indices of additional psychometric properties. These will be more complicated and more focused on their use in psychometric analysis. But even in this very simple case, we can see the important relationship between test difficulty and *test validity*. Not yet defining what "validity" is or should be, let's rely on common sense, saying that validity is the theoretical, as well as the practical usefulness of a test, to diagnose or measure the desired behavioral attribute. Perhaps it is intuitively clear that if test difficulty is either too high or too low, its score is less helpful for distinguishing between people within a given population. Either all measured subjects would be at the "floor" score level, or respectively, close to the "ceiling" score level.

Now, assume for a while that the practical purpose of a given test is just to distinguish between acceptable and non-acceptable subjects. As common sense would suggest, the best and most appropriate difficulty level would be if the test's cutting point were at "fifty/fifty" or the 50% level, having the scale cutting point of the test calibrated for the given population. If the test score is quantitative, such a test should be calibrated so that the overall conditions determining the task difficulty are adapted so that the cutting point of the score is located at the mean of the population. In the example of the *V-test*, we can find such a modification by instructing respondents from a population of very old and undereducated persons to produce words starting with the letter *P*. This would certainly be easier for this most frequent first letter in English than to produce words starting with *X* or *Q*, for example. Also, we could use a time interval longer than one minute to make a test more accessible. Such manipulation can yield a version of the original test that will have its mean satisfactorily close to the cutting point that had been established by experts, say for diagnosing a "normal" level of performance.

If a test takes on the "accomplished/failed" or "good/wrong" or similar dichotomized scores, then its psychometric index of difficulty is directly given by the corresponding percentage of persons who passed or failed. Such tests usually are only parts of a battery or inventory of tests, and traditional psychometric terminology calls them *test items* or simply *items*. Scores of such items take on only two possible values, therefore those are often called *binary* or *dichotomous items*, with values $y = 1$ and/or $y = 0$ for coding scores, usually by 1 for "right" and 0 for "wrong." Although this kind of assigning numbers seems to be an arbitrary symbolization, it still has an important property. It could be called *quasi-quantification*. Such an artificial random variable created by this (0, 1) coding of values is also well known as a *dummy* variable. It has certain quasi-quantified properties since in certain mathematical models it allows us to treat such tests and items *as if* they were measurable interval-metric quantities.

One of those mathematical methods is the *mean of a binary item*: $\overline{Y} = p$. The mean is now identical to the *proportion p* of unities "1" in the set of *N* subjects who successfully completed the test task. Since $\sum Y_i$ is the number of "1" values, when it is divided by the sample size *N*, it produces a proportion, $p = \frac{\sum Y_i}{N}$. Correspondingly, $q = 1 - p$ is the proportion of those who failed, and it follows that $p + q = 1$. Of course, the higher proportion of those who passed successfully, i.e. the higher the *p*, the easier the item is for a given population. But, strangely, perhaps due to a century-long tradition, so-called "item analysis" still calls *p* "the *difficulty* of a test item." We must just passively accept this illogical name, which by common sense should rather be called item "easiness," with an excuse that it is simply a technical term and a definition. Some people call *p* the "index of difficulty," perhaps an index with an opposite direction of meaning, but perhaps we could avoid the conflict by using intuition.

1.3.4 (§11) Intra-group and Inter-individual Variability—*Dispersion* of Scores

A more detailed look at Figure 1.7 suggests that the central tendency information does not tell the whole story of how the head circumferences of fathers compare with those of their sons. The overall level of the attribute and location of the group on the scale is not enough. Not only is the midpoint of the head circumference distribution for sons slightly higher than the midpoint for the fathers' distribution, but the sons' data distribution also has a somewhat larger spread, and a slightly larger dispersion.

Perhaps the two groups, the fathers and the sons, could be almost of the same mean level but still differ in their spread, that is, their variability. The first intuitive look, then, could be at the maximum and minimum values in each group. This is also visually apparent in Figure 1.7.

Returning to the two histograms we can use the formula and calculate the range for the group of fathers and the range for the group of sons in that figure, to verify numerically that the sons do have greater variability in head circumferences. The range for the fathers is:

$$Ra_f = max_f - min_f = 27'' - 19'' = 8''$$

And the range for sons is:

$$Ra_s = max_s - min_s = 29'' - 19'' = 10''$$

This contrast is also visually apparent in Figure 1.7.

The range $Ra(x)$ is a characteristic of variability, while the median characterizes location. Although each represents a different aspect of the distributions, they both still have a certain mathematical property in common. Like the median, the range is insensitive to small changes in individual values, until they jump over its border. However, the range can be overly sensitive to an extreme value of an atypical individual in the group whose value is outlying far to the side of the group. Such an *outlier* causes the information about variability to be distorted. Therefore, a better and "fairer" characteristic would be based on similar mathematical properties to those of the arithmetic mean so that even a very small change of any individual value influences the overall characteristic.

The idea that the larger the overall deviations of values from their center, the larger the dispersion, suggests that the arithmetic *mean of the deviations* can be such a "fair" characteristic. If the deviation of an individual subject's value x_i from the group mean is defined by $x_i = X_i - \overline{X}$, then a short-circuit conjecture could suggest that the

$$\text{"Mean Deviation"} \quad MD = \frac{\sum x_i}{N}$$

could be the characteristic we are looking for. But don't forget the mutual counterbalancing of the left (negative) and right (positive) deviations around the mean to keep the scale "lever" in equilibrium. Therefore, $MD = 0$ for any data, since $\Sigma d_i = 0$ for any set of data. This is a good check that the mean has been calculated correctly. Thus, *MD* is not of any particular use as an index of variability. A solution could be to take the mean of the absolute value of the deviations to produce the

$$\text{"Mean Absolute Deviation"} \quad MAD = \frac{\sum |x_i|}{N}.$$

Although this statistical characteristic is not really "mad," it is not "cool" either. It is not used much despite the fact that originally it had been suggested by the great French mathematician Pierre-Simon Laplace (1749–1827). Instead, the problem has been solved by a famous British statistician of the 20th century, Sir Ronald Fisher the inventor of variance. Fisher's approach is not to take absolute values, but to avoid the negative values by computing their squares. Thus, instead of

MAD the better index of spread is the MSD, based upon a sum of squared deviations, a special case of the Bravais (1846) "product sum," or "sum of products."

$$\text{"Mean Squared Deviation"} \quad MSD = S_x^2 = \frac{\sum {x_i}^2}{N}.$$

It has become one of the most basic statistical concepts, and it is known under the name of variance. The variance of variable X is written out in longhand as:

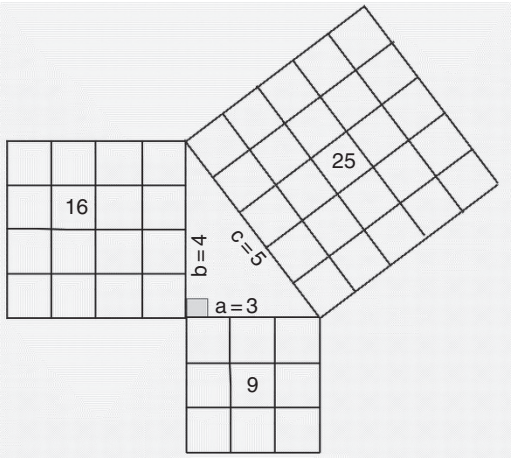
$$S_x^2 = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{N}$$

At the first glance, $MSD = S_x^2$ seems more complicated than *MAD*, since it is a number expressed in squares of the original measurement units. This last inconvenience can be quite intuitively corrected—after computing squared deviations and averaging them we can return to the original units by taking the square root of the variance to get the standard deviation, also called SD.

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{N}}$$

Galton first proposed the standard deviation in the late 1860s, just a few years after Bravais (1846) published the product sum. Note that S, by the arithmetic operations taken—and intuitively, too—is also a kind of "average" deviation like *MAD*.

Why was MAD dropped in favor of variance? At least two reasons can be given. One is that Gauss, the same mathematician who discovered the Gaussian distribution (the normal curve), and also the least squares principle on which regression is based, had the foresight to realize the profound advantage of using *squared* units for his least squares principle on which regression is based, which places it squarely in the mainstream of Euclidean geometry and the Pythagorean Theorem. Squared units are additive. Note in the classic 3–4–5 triangle at the right, that the sides are not additive ($3 + 4 \neq 5$), but their squares are ($9 + 16 = 25$). Sir Ronald Fisher made squared deviations the foundation of his variance statistic, published in 1918. It has become one of the foundation pieces of statistical theory. Regression and multiple regression (Appendices 4 and 5) also embody the additivity principle, in which the total variation in Y can be decomposed into predicted variation and error variation. Also, in CTT the total variance in a test is decomposable into true-score variance and error variance.



The second reason is that the standard deviation (SD), the square root of variance is a number, a statistic (that is, a sample index), with the special property of becoming[14] a parameter (a population index), which in some data distributions found in nature enables us to compute the relative frequency of the data from a normal curve table

14 In the first part of Section §13, in Table 1.3 and the accompanying text, we will explain the slight modification of the variance & standard deviation to transform them from sample statistics into unbiased population parameters.

and thus estimate probabilities of occurrence. As well, it is the parameter that allows us to estimate the attributes of a whole population if we know the SD in just a small random subsample. Thus, it makes it possible to establish the probable error of estimation for a whole population. Also, it enables us to compute the relative position of a subject within his or her sample on a standardized test, or even within a specific population. It also can be used to establish norms and standards of clinical normality, estimate the probability that a subject fulfills the norms, and so forth, as will be shown in Sections §12 and §13.

1.3.5 (§12) A Small Numerical Example, Finally

Didactic doctrine says that a theoretical presentation should be accompanied by examples that are illustrative and real. Didactic experience teaches that "real" examples are often too complicated to be illustrative. So, please accept the next artificial but simple example, patterned after the anthropomorphic measures of Galton, with the apology of the authors.[15] Imagine that we have selected $N = 6$ fathers and their sons from among the group of $N = 30$ father/son pairs whose data are shown in Figure 1.7. The head circumference values for the six selected fathers are shown in Table 1.1, together with the calculation of the mean of the head circumferences, the deviations of the circumferences from that mean, and the sum of squared deviations of the circumferences from the mean ($\sum x_i^2$).

The variance of the circumferences is then calculated by dividing this sum of squared deviations, 24, by N which gives 4 as the variance of the six fathers' head circumference measurements.

Table 1.1 Fictional head circumferences (X_i) of six fathers showing calculation of the sample total (T), the sample mean ($\overline{X}$), deviations from the mean (x_i), squared deviations from the mean, and the sum of the squared deviations from the mean ($\sum x_i^2$).

Family name	X_i	x_i	x_i^2
Abbott	26	3	9
Bailey	25	2	4
Cardon	23	0	0
Davis	22	−1	1
Evans	22	−1	1
Flanders	20	−3	9
$T = \sum x_i = 138$			$\sum x_i^2 = 24$
$\overline{X} = \frac{\sum X_i}{N} = \frac{138}{6} = 23$			

15 We have used small samples here to make the demonstration of the calculations for the descriptive statistics simpler and easier to follow. In actual practice, one would usually be interested in inferring from the sample to some population, so-called inferential statistics as introduced in the next section, and would expect to utilize substantially larger samples, selected at random, if possible, to ensure an adequate estimation of the distributional properties of the population.

$$S^2_{fathers} = \frac{\sum x_i^2}{N} = \frac{24}{6} = 4$$

Taking the square root of this, we get a standard deviation of 2, which is interpretable in inches. The sample standard deviation of the six fathers' head circumferences is therefore 2 inches.

$$S_{fathers} = \sqrt{4} = 2$$

Using the Excel command "=**normsdist(1)-.5**,"[16] it can be determined that, as shown in Figure 1.8, the proportion of total area under the normal curve expected to fall between the mean and one standard deviation unit above the mean is 0.3413, to four decimal places. Since the normal curve is symmetrical, the area of the normal curve between the mean and one standard deviation below the mean is also 0.3413. The sum of the two areas is 0.6826. In other words, if a data set is normally distributed, about 68% of the scores will fall between +1 and −1 standard deviations from the mean. The mean head circumference for this small sample of fathers is 23 inches, and we have just found the standard deviation to be 2 inches. So, we would expect about two-thirds of the scores to be located between 21 and 25 inches (23 inches ± 2 inches, that is, the mean plus or minus the standard deviation).[17]

With the command "=**normsdist(2)-.5**," it can also be seen that the area between the mean and two standard deviation units above the mean is 0.47725. Summing that area with a 0.47725 area from the left half of the distribution, the percent of scores falling between ± two standard deviation units from the mean is a little over 95% (95.45% to be more precise), so a *standard normal curve*, that

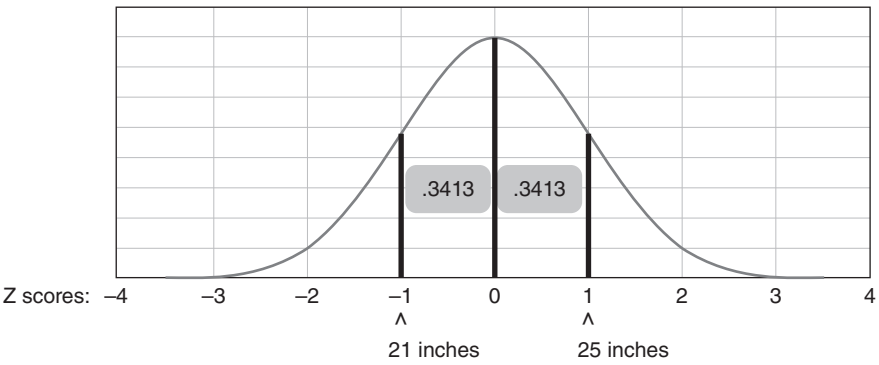


Figure 1.8 The area of the standard normal curve distribution between +1 and −1 standard deviation units is 0.3143 + 0.3413 = 6826, or about 68% of the area. Given that the six fathers have a mean head circumference of 23 inches with a standard deviation of 2, that means that we should expect about 2/3 of the fathers to be within this range. In Table 1.1, the middle four of the six fathers have circumference values of 22, 22, 23, and 25.

16 The Excel command "=normsdist(1)" requests the proportion of area under the normal curve that is below a *Z* score of 1.00, and the answer is 0.8413. But we want the area between a *Z* score of 1.00 and a *Z* score of 0.00 (the mean) so we append "−0.5" to the command to subtract the area in the lower half of the curve from 0.8413 to get 0.3413. For many years statistics books and psychometrics books have included statistical tables such as a table of areas under the standard normal distribution. Given the ready availability of statistical table functions such as this in spreadsheets such as Excel, the need for such tables is somewhat reduced.

17 With simple demonstration data such as this we would not expect the probabilities from the normal curve to be anything but very approximate. It is shown here to illustrate the process of using statistical tools and probability tables with simple data as a guide in preparation for use with research data.

Table 1.2 Fictional head circumferences (Y_i) of the sons of the six fathers in Table 1.1, showing the calculation of the sample total (T), the sample mean ($\overline{Y}$), deviations from the mean (y_i), squared deviations from the mean, and the sum of the squared deviations from the mean ($\sum y_i^2$).

Family name	Y_i	y_i	y_i^2
Abbott	26.4	2.95	8.703
Bailey	24.3	0.85	0.723
Cardon	23.5	0.05	0.003
Davis	24.4	0.95	0.903
Evans	19.6	−3.85	14.823
Flanders	22.5	−0.95	0.903
$T = \sum Y_i = 140.7$			$\sum y_i^2 = 26.055$
$\overline{Y} = \frac{\sum Y_i}{N} = \frac{140.7}{6} = 23.45$			

is, a normal curve with Z-score units for its horizontal axis, has approximately 68% of its area between one standard deviation above and below the mean and approximately 95% of its area between two standard deviations above and below the mean.

Consider now comparable measurements and calculations for the sons of these $N = 6$ fathers, for a comparison similar to that used by Galton. The head size measurements of the sons of these six fathers are presented in Table 1.2 together with their mean, their deviation scores from the mean, the squared deviation scores, and finally the sum of squared deviation scores. The variance and standard deviation are calculated for the head size measurements of the sons, using the same mathematical process just shown for the fathers' data. Using the sum of squared deviation scores of 26.055 from Table 1.2 the variance of the sons' head circumference values is calculated to be 4.34 and the standard deviation is the square root of 4.34, which is 2.08 inches:

$$S^2_{sons} = \frac{\sum y_i^2}{N} = \frac{26.055}{6} = 4.34$$

In Section §4 we discussed Galton comparing means and standard deviations of head circumferences for fathers and sons, and his development of the reversion or regression line to predict sons' head circumferences from that of their fathers. In this section, we have shown the statistical process of how those comparisons can be made. We have shown that not only do the sons have slightly larger mean head circumferences than their fathers for this small fictional example, but they also have a somewhat larger standard deviation. In the final section, Section §13, we will introduce inferential statistics in connection with the *Central Limit Theorem*, and demonstrate how statistical inference with sample sizes greater than 30 can give aggregated precision reminiscent of the double-dice tosses shown in Figure 1.6.

1.3.6 (§13) Does the Larger Head Circumference in the Sample of Sons Than Fathers Hold for the Whole Population? A Preliminary Sidestep to Statistical Inference

This question may be answered if we use a higher type of statistical reasoning, beyond *descriptive* statistics, that is, beyond those statistics that merely describe subjects within a particular sample. But this is possible only if some important statistical conditions have been met, enabling us to move

from descriptive statistics to *inferential statistics*. This branch of statistics offers methods for making an inference, a generalization, from a given representative *sample mean* and *sample standard deviation* to the corresponding characteristics of the whole population, but there is a caveat. Statistical inference works *only if the mathematical assumptions hold in reality*.

The first of these mathematical assumptions is a "randomly selected sample drawn from a clearly defined population." Obviously, a card register or a computer database of patients in a clinical psychologist's office is not comparable to a randomized representative sample of a city or a county. And, of course, the statistical meaning of the target "population" need not have a geographic or demographic reference at all. A *random sample* from a population is defined as *one in which each person or element in the population has an equal chance of being selected*.

An important reason why this topic is presented here and now, so early in the text, is that many students ask why standard deviation is sometimes symbolized by S or sometimes by σ, and why clinicians and researchers speak about "how many sigmas" from the mean is an acceptable limit of normality for a score. Moreover, some ask why the formula for calculating standard deviation is presented sometimes with division by N and elsewhere with division by $N-1$. The answers to questions of this kind require an understanding of inferential reasoning.

Suppose that we would be satisfied with just knowing how to compute a "one-valued" estimate of the true population mean or the true population variance, an estimate without additional information, say about its possible error. This is called a *point estimate*. In statistics, it is usual to use Greek letters to denote population characteristics, so we will use μ to represent a point estimate of the population mean and σ^2 to represent a point estimate of the population variance. From a statistical point of view, we can say that in a randomized sample, the mean $\overline{X}$ is the "best" *point estimate* of the population mean μ. Being the "best" point estimate requires that in a large series of many repeated random samples, the $\overline{X}$ will not systematically under-estimate nor over-estimate μ, the true population mean value. We can say, therefore, that the sample mean $\overline{X}$ is an *unbiased* estimator of the true population mean, μ.

But this unbiasedness does *not* hold for the descriptive sample variance S^2 ("s squared") as an estimate of the population variance σ^2 ("sigma squared"). The formula for the descriptive sample variance S^2 is given in Equation 1.1.

$$S^2 = \frac{\sum x_i^2}{N} = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{N} \qquad \textit{(Equation 1.1, descriptive sample variance)}$$

It is nothing more than the average of the squared deviation scores, that is the sum of the squared deviations of scores from their mean ($\sum x_i^2$) divided by N. Since the standard deviation is the square root of variance, the unbiasedness also does not hold for the sample standard deviation.

If we were to use an ordinary descriptive sample variance such as this as the point estimate of the population variance, it would systematically underestimate the true population variance σ^2 by "one Nth" of its value. That is, if the true population variance were 12.0 and if the sample size were $N = 100$, the average of many such estimates would be about 11.88, underestimating the true population variance by 0.12, one one-hundredth of the value of σ^2. However, if the sample size were only 10, the average estimate of the variance over many trials would be about 10.80, underestimating the variance by 1.20, one-tenth of its true value. Obviously, estimation bias is much stronger with small sample sizes. The evidence for this sampling principle is presented at the end of this final section (Section §13) in Figure 1.15 and its accompanying text.

Fortunately, the demonstration given in Figure 1.15 not only verifies this principle, it also informs us as to what the magnitude of the systematic error is and how to correct for it. The correction is

simple. In the computation of the sample estimate of the population variance, instead of dividing the sum of squared deviations by the sample size N, we divide it by the *degrees of freedom* $N-1$,[18] as shown in Equation 1.2. The square root of this unbiased variance yields the unbiased standard deviation as shown in Equation 1.3. This is the correct formula to use for inferential estimation of the population standard deviation, the formula that will return an *unbiased point estimate of the true population standard deviation*.

$$\hat{\sigma}^2 = \frac{\sum {x_i}^2}{N-1} = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{N-1} \qquad (\textit{Equation 1.2, unbiased estimate of the population variance})$$

$$\hat{\sigma} = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{N-1}} \qquad (\textit{Equation 1.3, unbiased estimate of the population standard deviation})$$

Notice the *hat* (^) over the sigma in Equations 1.2 and 1.3. It is usual to place a "hat" over a statistical estimate to indicate that it is *statistically* "*good*," that is, "plausible" in one of the meanings used in statistical theory. Here, it happens to be a good inferential estimate, an unbiased estimate of a population characteristic on the basis of sample data. When the sample size is large, the degree of freedom adjustment makes little difference. But for small samples, the difference can be quite sizeable. We encounter another meaning of "good estimation" in the context of using a regression analysis to predict a Y variable from X variables, as shown in Appendices 4 and 5. Since the regression method fulfills certain requirements for statistical optimality, namely the "least squares" criterion, the hat ^ symbol is used for regression estimates as well.

Table 1.3 presents both the sample descriptive statistics and also the unbiased point estimates of the population parameters for the fictional sample data of six father-son pairs from Section §12. With the small N of only six fathers and six sons, this table reveals the somewhat sizeable disparity between the descriptive sample standard deviation *statistics* on the left and the unbiased point estimates of the population standard deviation *parameters*[19] on the right. Since the sample mean is an

Table 1.3 Descriptive sample standard deviations and means for the six father/son pairs from Tables 1.1 and 1.2 compared to the corresponding unbiased point estimates of the population values.

Sample descriptive statistics			**Unbiased point estimates of population parameters, based on recalculation of variances using degrees of freedom**		
	Means	**Standard deviations**		**Means**	**Standard deviations**
Fathers	$\overline{X} = 23.00$	$S_x = 2.00$	Fathers	$\overline{X} = 23.00$	$\hat{\sigma}_x = 2.19$
Sons	$\overline{Y} = 23.45$	$S_Y = 2.08$	Sons	$\overline{Y} = 23.45$	$\hat{\sigma}_y = 2.28$

18 In beginning to analyze a sample with $N=6$ observations, all six values are free to vary during the calculation of the mean. The mean retains all 6 degrees of freedom. However, in the process of calculating a variance, the sum of squared deviations from the mean cannot be accomplished without first calculating the mean, which uses up one of the 6 degrees of freedom. Therefore, the number of degrees of freedom for calculating the variance is $N-1$, or 5 in this example.

19 Notice that a sample summary value, such as a mean $\overline{X}$, is referred to as a *statistic* and is expressed in ordinary Arabic letters, while the corresponding summary population value, such as the population mean μ, is referred to as a *parameter* and is expressed in Greek letters. One could say that statistic is to parameter as sample is to population.

unbiased estimate of the population mean, the means for the six fathers and the six sons on the left side of this table are the same as those on the right side. However, the standard deviations differ substantially. The computation of the standard deviations on the left follows the definition that variance is the arithmetic average of the squared deviation scores. On the other hand, for the unbiased point estimates of the population variances, on the right, it is essential to divide the sum of squared deviations by $N - 1$ rather than N. The square root of this unbiased variance is an important statistical achievement: an *unbiased point estimate of the unknown population standard deviation*. However, the point estimate could also be biased by inadequate sampling. Remember that it will only be an unbiased point estimate if the measurements were performed on a randomly selected sample.

A numerical index calculated from sample data in this statistically optimal way, by random sampling and by using the correct formula, is referred to as a valid and proper *statistical estimate*. There is one additional condition. It would be wise to obtain a sufficiently large sample to obtain a stable estimate of the mean. A small sample size was used here as part of a simplified and more transparent demonstration and to emphasize the strong effects of bias in a small sample. In an actual inferential estimate of population variance, a larger sample size would be called for. How large would depend on the purposes for the estimate and the level of precision needed.

If a larger sample of pairs of fathers and sons had been obtained from a properly randomized sample from a population, and if unbiased point estimates of the standard deviations of fathers and sons had been obtained in the manner described, then we could answer "yes" to the question in the title of this section. Under these conditions, it would be possible to determine whether the slightly larger head circumference for sons than fathers would hold for the whole population. Looking back to Section §11 where the logic of variance was explained, it should be emphasized that another reason for choosing the variance and the standard deviation in preference to the *MAD* statistical index as measures of spread is that no statistical theory is available for modifying *MAD* into a statistic that could enable optimal estimation of the corresponding population value in this way.

1.3.6.1 An Estimate of the Probable Error of a Point Estimate

As valuable as it is to have unbiased point estimates of population means, variances, and standard deviations, it is of substantially more value if it can also be combined with an estimate of the accuracy of those point estimates, that is, an estimate of probable error. This is true in psychometrics and also in the more general discipline of metrology, the scientific study of measurement. Brown (2021), in his article "Measuring measurement—What is metrology and why does it matter?," mentions the measurement of "uncertainty," another word for probable error, as one of the primary concerns of metrology. An evaluation of the uncertainty present in one's point estimate is essential to good measurement. To learn this process of estimating uncertainty, the probable error of one's measurements, it will be necessary to now take the "preliminary sidestep into inferential statistics" referred to in the subtitle of this section.

As a thought experiment, suppose we create a large collection of unbiased point estimates of the population mean taken from many random samples of a given population to evaluate how well they agree with one another. This collection of sample means is referred to as an *SDM*. The standard deviation of this SDM is called the *standard error of the mean* $\left(\hat{\sigma}_{\overline{X}}\right)$. It would indicate how large the spread is among all of the sample means. A large spread would be indicated by a large *standard error of the mean* $\left(\hat{\sigma}_{\overline{X}}\right)$, and a small spread would be reflected in a small standard error of the mean.

However, thanks to one of the principles of inference, the tedious task of assembling a large collection of samples from which to calculate $\hat{\sigma}_{\overline{X}}$ is not necessary. This short-cut principle is expressed

in Equation 1.4, which indicates that the variance of means is one *N*th as large as the variance of the corresponding raw scores.[20]

$$\hat{\sigma}^2_{\overline{X}} = \frac{\hat{\sigma}^2}{N} \qquad (\textit{Equation 1.4, the formula for the squared standard error of the mean})$$

This indicates that if one has an unbiased estimate of the population variance, Equation 1.4 can adequately estimate from that variance the *squared standard error of the mean*, or in other words the variance of the *SDM*. Taking the square root of this equation yields the *standard error of the mean* (Equation 1.5), which is the standard deviation of the SDM. This gives an unbiased estimate of the *probable error* for an estimated population mean.

$$\hat{\sigma}_{\overline{X}} = \frac{\hat{\sigma}}{\sqrt{N}} \qquad (\textit{Equation 1.5, the formula for the standard error of the mean})$$

For the fictional data in Table 1.3, the mean of the head circumferences for the six sons was found to be $\overline{Y} = 23.45$. The standard error for that mean is calculated using Equation 1.5.

$$\hat{\sigma}_{\overline{Y}} = \frac{\hat{\sigma}_y}{\sqrt{N}} = \frac{2.28}{\sqrt{6}} = \frac{2.28}{2.449} = 0.931 \qquad (\textit{Equation 1.5a}^{21})$$

This is an unacceptably large standard error of the mean, but that is because we have such a small *N* of 6 for this simple demonstration data. Suppose that the *N*′ for the sample of these father/son pairs were 80 rather than 6. We now have a larger sample size for *Y*, denoted as *Y*′, but having the same means and the same standard deviation as given in Table 1.3. The standard error for the mean of the sons is now of a more reasonable size.

$$\hat{\sigma}_{\overline{Y}'} = \frac{\hat{\sigma}_{y'}}{\sqrt{N'}} = \frac{2.28}{\sqrt{80}} = \frac{2.28}{8.944} = 0.255$$

To clarify the measurement uncertainty for this estimated mean as related to its standard error we will use a confidence interval. There are two steps in creating a confidence interval. The first is to decide on the confidence level chosen for the interval, and the second is to calculate the upper and lower score values that are the boundaries of that confidence interval. Perhaps we would like to be 95% confident that our population mean is within the confidence band. Using the Excel command "**=normsinv(.975)**,"[22] it is determined that a *Z* score of 1.96 has an area below it of 0.975 which corresponds to an area of 0.025 above it on the right side of the standard normal curve. Similarly, a *Z* score of −1.96 has an area of 0.025 below it on the left side. The 95% confidence region therefore lies

20 One would expect the variance of means to be less than the variance of raw scores given the effects of averaging, but the surprising thing is that the relationship is so simple. If one were to average raw data in randomly created groups of five, these means would be expected to have a variance that is one-fifth that of the raw scores. If one averages in groups of ten, the variance of the means would be one-tenth of the variance of raw scores. In the language of statistical inference this quantity is referred as the *variance of the sampling distribution of means*, but it is also called, and is equivalent to, the *squared standard error of the mean*. This equation is also an aspect of the *Central Limit Theorem* introduced in Section §7. A demonstration of its validity is found in Figure 1.14 and accompanying text at the end of this section.

21 Notice that this is still Equation 1.5, but now applied to variable *Y*. When introducing the formula, it was shown for variable *X*. In other words, in its first appearance the symbol for the formula was $\hat{\sigma}_{\overline{X}}$, which is read as "the standard error of the $\overline{X}$ mean." In the second appearance the symbol was $\hat{\sigma}_{\overline{Y}}$, which is read as "the standard error of the $\overline{Y}$ mean"—the same formula, but applied to a different variable.

22 The Excel command "normsinv(.975)" requests the *Z* score value that has an area below it of 0.975, which returns a *Z* value of 1.96.

between $Z = -1.96$ and $Z = 1.96$ on the standard normal curve. The confidence interval now needs to be expressed in raw score units rather than Z score units. Equation 1.6 will create the desired confidence interval.

$$CI_{95} = \overline{X} \pm Z_{0.025}\,\hat{\sigma}_{\overline{X}} \qquad (\textit{Equation 1.6, confidence interval for a single mean})$$

Applying this equation to variable Y' with a sample size of 80 and the same mean and standard deviation given for the sons in Table 1.3, we obtain a confidence interval of 23.45 ± 0.50.

$$CI_{95} = \overline{Y} \pm Z_{0.025} \quad \hat{\sigma}_{\overline{Y}} = 23.45 \pm (1.96)(0.255) = 23.45 \pm 0.50$$

The lower limit of this confidence interval is equal to $23.45 - 0.50 = 22.95$. The upper limit is $23.45 + 0.50 = 23.95$. The 95% confidence interval for the population mean of the head circumferences of sons therefore extends from 22.95 to 23.45. Since the mean head circumference of fathers was found to be 23.00, and lies within the confidence band for sons, we conclude that we do not have sufficient evidence to say that the larger head circumference in the sample of sons than fathers holds for the whole population.

In the past several pages, some of the principles of statistical inference have been introduced: how to obtain a point estimate of the population mean, how to obtain an unbiased estimate of the population variance that can be used to derive the standard error (that is the probable error) of that population mean, and how to combine these two statistics to construct a descriptive and interpretable confidence interval within which one can be 95% confident the true population mean lies. These are some of the uses and principles of stochastic (probabilistic) modeling, one of the three preconditions specified in the last four paragraphs of Section §5 for creating "bona fide research data".

In the following three subsections, additional examples of stochastic models will be presented to summarize and clarify the central fundamental principles of statistical inference. These stochastic models will function like brief "parables" that demonstrate principles in a mathematically simpler form using binary data. In the fourth and final remaining subsection, we will expand from binary to multinary data, using the multinomial distribution, to illustrate the principles of inference applied to interval/ratio data.

1.3.6.2 Classical Probability Theory and the Binomial Distribution: Creating a Simple Stochastic Model

Statistical inference is based upon classical probability theory, the mathematical creation of a coterie of "natural philosophers" during the Age of Enlightenment. It began with the pioneering work of Jakob Bernouli in the 17th century and extended into the 18th and 19th centuries with the work of theorists such as Pierre-Simon Laplace (Daston, 1995). Though "probability theory" sounds daunting, the fundamental principles can be made quite understandable, as stated by Laplace in the introduction to his 1812 book, *Théorie Analytique des Probabilités*,

> The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which ofttimes they are unable to account.

Classical probability theory and inferential statistics are often taught using binary data, which serves as a simpler proxy for quantitative data. This works well. The mathematical procedures for means and variances are often the same for binary data as for fully quantitative interval/ratio data. For an illustrative example, suppose that surveys have been taken of citizen approval of a new

law adopted in a small town. One survey measures the level of approval using a 10-point scale. A second survey is conducted using binary "yes" or "no" answers. The first six persons to answer the quantitative survey give ratings of 5, 7, 3, 4, and 2. On the other hand, the first six persons to answer the binary second survey respond with "yes," "yes," "no," "yes," and "no," which responses are then "dummy coded" as 1, 1, 0, 1, 0. The mean of the quantitative survey is calculated as:

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{5+7+3+4+2}{5} = \frac{21}{5} = 4.2$$

The mean of the second survey, which is actually a proportion since it was calculated for binary data, is found to be:

$$p = \overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{1+1+0+1+0}{5} = \frac{3}{5} = 0.60$$

In both cases, the response values for the six persons are just summed and divided by N, the number of responses. Three out of the five responses on the second survey, or 60% of them, were "yes" responses. This corresponds to a proportion of 0.60. It is mathematically equivalent to the mean calculated in the usual way. It is the mean of a variable consisting of only ones and zeros. Simply stated, the proportion is a *special case of the mean*. It is the mean of a binary variable.

The descriptive sample variance of the quantitative survey is calculated as:

$$S^2_{quant} = \frac{\sum x_i^2}{N} = \frac{\sum \left(X_i - \overline{X}\right)^2}{N} = \frac{(0.8)^2 + (2.8)^2 + (-1.2)^2 + (-0.2)^2 + (-2.2)^2}{5} = \frac{14.8}{5} = 2.96$$

The descriptive sample variance of the second survey, the binary one, can be calculated in the same manner, simply the sum of the squared deviations of scores from their mean, divided by N:

$$S^2_{binary} = \frac{\sum x_i^2}{N} = \frac{\sum \left(X_i - \overline{X}\right)^2}{N} = \frac{(0.4)^2 + (0.4)^2 + (-0.6)^2 + (0.4)^2 + (-0.6)^2}{5} = \frac{1.2}{5} = 0.24$$

There is also, however, a much simpler way to calculate the variance of the binary variable. We just calculated the proportion for this binary variable, that is its mean as $p = \overline{X}$, and found it to be 0.60. If q is defined as "one minus p," the variance of the binary variable may be calculated simply as pq, the product of p and q as shown in Equation 1.7.

$$S^2_{binary} = pq = p(1-p) = (0.6)(0.4) = 0.24 \qquad (\textit{Equation 1.7, the variance of a proportion})$$

Surprisingly, this variance is not only the descriptive sample variance of the binary data of the second survey, it is also the unbiased point estimate of the corresponding variance of the binary data in the population.[23] The square root of this variance is therefore the unbiased standard deviation of the binary data in the population. In other words, the formula $\hat{\sigma} = \sqrt{pq}$ is used to obtain an *unbiased point estimate of the true population standard deviation of binary data*, which for our example data from the yes/no survey is found to be: $\hat{\sigma} = \sqrt{pq} = \sqrt{(0.6)(0.4)} = \sqrt{0.24} = 0.490$.

23 With data that is continuous and quantitative, the variance can only be calculated after the mean has first been calculated, since the mean is first needed to obtain the deviations from the mean for the "sum of squared deviations" used in calculating the variance. This uses up one degree of freedom. However, as demonstrated above, for binary data the variance can be calculated directly from the proportion without obtaining squared deviations from the mean, so there is no loss of degrees of freedom. It follows that, for binary data, the formula $\hat{\sigma}^2 = pq = p(1-p)$ gives an unbiased point estimate of the population variance, and $\hat{\sigma} = \sqrt{pq} = \sqrt{1-p}$ gives an unbiased point estimate of the true population standard deviation.

This estimate of the standard deviation of the binary data in the population can in turn be used to obtain an estimate of probable error, that is, of the *standard error of the proportion*, by dividing *pq* by *N* before taking the square root, as shown in Equation 1.8.

$$\hat{\sigma}_p = \sqrt{\frac{pq}{N}} \qquad (Equation\ 1.8,\ standard\ error\ of\ the\ proportion)$$

For this small example, the *standard error of the proportion* is calculated to be:

$$\hat{\sigma}_p = \sqrt{\frac{pq}{N}} = \sqrt{\frac{(0.6)(0.4)}{N}} = \sqrt{\frac{0.24}{5}} = \sqrt{0.048} = 0.219$$

Additional fundamental principles of probability theory will now be demonstrated using the *binomial distribution* to create sampling distributions of means (or more precisely of proportions) for binary data. Consider, for example, the common act of coin tossing for our first stochastic model. Coin tossing has the advantage of being familiar, simple, concrete, and memorable.

The Multiplication Rule for Independent Joint Events If a fair coin[24] is tossed three times, one of eight things can happen. It can come up with three heads in a row (HHH) as shown in the first branch of Figure 1.9. Or, it could come up with two heads in one of three ways (in the order of HHT as shown in the second branch, or HTH as in the third, or THH as in the fifth). Or, it could come up with a single head in one of three ways (in the order HTT as shown in the fourth branch, THT as in the sixth, or TTH as in the seventh). Or, finally, it could come up with zero heads in one way (TTT), as shown in the eighth branch of Figure 1.9. To the right of the branching diagram is a pattern listing in the same order from top to bottom as the branching diagram, but with probability calculations

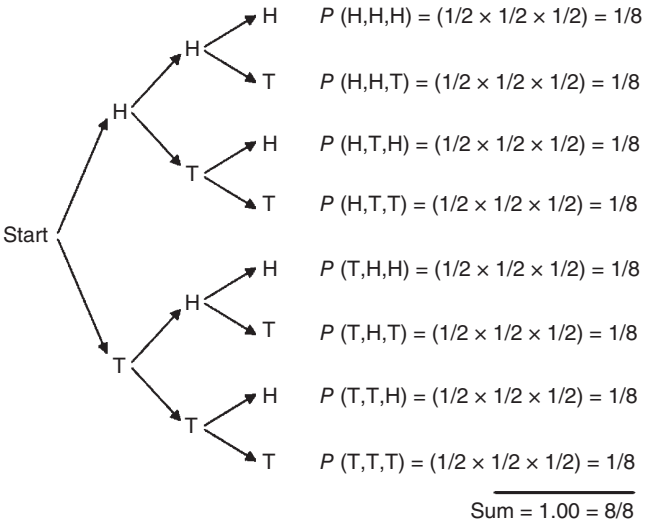


Figure 1.9 Branching diagram showing the eight possible patterns of heads and tails created by "fair coin" tossing ($p = 1/2$) in these-toss trails, and the calculation of each their respective probabilities.

24 A "fair coin" is a coin that comes up heads half of the time and tails half of the time. That is, over the long haul the expected value of its proportion of times coming up heads, its mean number of heads is 0.50, when "heads" is dummy coded as "1" and tails is dummy coded as "0."

added for each of the eight outcome branches. Since this is a fair coin, the probability for each head is always ½ and the probability for each tail is always ½ also. The probability for each branch of the diagram is just the product of individual probabilities for each of the three events (H or T) in sequence, which comes out to 1/8 in all eight patterns of the $p = ½$ model. This is known as *the multiplication rule for independent*[25] *joint events*, Equation 1.9. It is used for "AND" probabilities. It is written as:

$$P(A \text{ AND } B) = P(A) \times P(B) \qquad (\textit{Equation 1.9, the multiplication rule})$$

In set theory notation, "AND" probabilities are referred to as "intersection" probabilities. The multiplication rule is written in set notation as "the probability of A intersect B is equal to the probability of A multiplied by the probability of B":

$$P(A \cap B) = P(A) \times P(B)$$

The multiplication rule can, of course, also be generalized to three or more independent joint events as shown in this set-theoretic version of the equation:

$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C) \qquad (\textit{Equation 1.9 expanded for three joint events}^{26})$$

As Laplace stated, probabilities are nothing but "common sense reduced to calculus," or in other words common sense logic expressed with numerical calculations. From a logical standpoint, the probability of getting a single head in one toss is one-half (½), but the probability of getting two heads in two tosses is half of that half, or one-fourth ($½ \times ½ = ¼$), as can be seen by examining the last two tosses in the first four outcome patterns of Figure 1.9 (HH, HT, TH, and TT, only one of which is two heads). The probability of getting three heads in three tosses is half-of-a-half-of-a-half, or one-eighth ($½ \times ½ \times ½ = ⅛$). Of course, since this example in Figure 1.9 gives results for tossing a fair coin, the probability is always 0.50 whether heads or tails, and the probability of TTH or any other one of the eight outcome patterns is always one eighth ($½ \times ½ \times ½ = ⅛$) as shown in the individual calculations of Figure 1.9. This is the simple and unique branching diagram that results when $p = q$, and each of them is therefore equal to ½ (0.50).

It must also be noted that this is a probability model, so the outcomes are *mutually exclusive*, meaning that if the outcome is HHT it cannot also be THT or any other one of the remaining six specified outcomes of the eight branches. The outcomes are also *exhaustive*, meaning there are no other possibilities than the eight that are listed. Finally, the *probabilities must sum to 1.00*, which, of course, they do since there are eight possible outcomes specified, and each has a probability of one-eighth as shown in the right half of Figure 1.9. These two paragraphs are an example of the "common sense" part of Laplace's statement, the logic of probability. The following

25 Some joint events are *dependent joint events*, meaning that the first event occurring has an impact on the probability of the second event occurring. For our purposes, we will only consider independent joint events, that is joint events where the occurrence of the first event has no effect on the probability of the second event. The equation for the *general multiplication rule of probabilities* is $P(A\_\text{AND}\_B) = P(A) \times P(B|A)$, where $P(B|A)$ refers to "the probability of B given that A has occurred," in other words $P(B)$ can be changed by the prior occurrence of event A. For more information about the general rule, go to Khan Academy online or do a web search on "the general multiplication rule of probabilities."

26 When we ask the question "how many possible patterns are there in three (or four, or five, etc.) tosses of a coin?" we are thinking of this as a "joint event" consisting of three, or four, or any number of combined single events. Equation 1.9 can therefore be expanded to any number of independent joint events in the same way we expanded it from two-toss joint events to three-toss joint events.

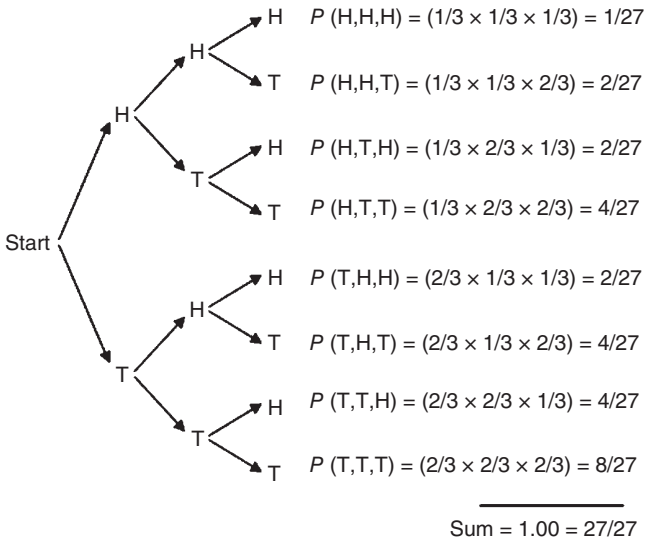


Figure 1.10 Branching diagram showing the eight possible patterns of heads and tails created by "weighted coin" tossing ($p = 1/3$) in three-toss trails, and the calculation of each of their respective probabilities.

paragraphs and their accompanying figures and tables spell out the "reducing-it-to-calculus" part, that is expressing the logic as calculations.

Whereas the probability model in Figure 1.9 has equal coin-tossing probabilities for heads and tails, our second stochastic model in Figure 1.10 is more complex. It is a "weighted coin" model with probabilities of $p = ⅓ = 0.333$ and $q = ⅔ = 0.667$. Although these two stochastic models have similar structures, each is a separate and distinct probability model, because they have differing probabilities. Whereas the eight branches in Figure 1.9 "fair coin" branching diagram have the same probability (all are $P = 1/8$), those in Figure 1.10, the "weighted coin" model do not. Because it is more complex, the "weighted coin" branching diagram gives more insight into the nature of probability.

The Addition Rule for Mutually Exclusive Composite Events The probability for each branch of Figure 1.10 can be obtained using Equation 1.9. For example, the probability for the second branch down from the top (HHT) is calculated by defining $P(A) = 1/3$ in the equation as "the probability of H, a head, on the first toss," $P(B) = 1/3$ as "the probability of H on the second toss," and $P(C) = 2/3$ is defined as "the probability of T, a tail, on the third toss." As shown in Figure 1.10, the joint probability of this HHT sequence in the second branch is calculated to be the product of those three probabilities, which equals 2/27.

$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C) = P(H) \times P(H) \times P(T) = \left(\frac{1}{3}\right) \times \left(\frac{1}{3}\right) \times \left(\frac{2}{3}\right) = \frac{2}{27}$$

These same three probability values also appear (although with their order permuted) in the other two outcome patterns that are characterized by the description of "getting two heads in three tosses of a weighted coin," namely HTH, the third branch down; and THH in the fifth branch. It follows that they each also have a probability of 2/27. If one asks the question "what is the probability in the model of Figure 1.10 of getting exactly two heads in three tosses of a coin?," the answer must be "the combined sum of the probabilities of these three patterns, HHT, HTH, and THH,"

since all three fit this same description of "two heads in three tosses of a coin," and none of the patterns in the other five remaining branches does.

$$P(2 \text{ heads}) = P(\text{HHT}) + P(\text{HTH}) + P(\text{THH}) = \frac{2}{27} + \frac{2}{27} + \frac{2}{27} = 6/27$$

This is *the addition rule of mutually exclusive*[27] *composite events*.[28] It is used for "OR" probabilities, the probabilities that are additive. Equation 1.10 below gives the basic equation for the addition rule with two mutually exclusive events and Equation 1.10a is an expanded version for three such events.

$$P(A \text{ OR } B) = P(A) + P(B) \quad (\textit{Equation 1.10, the addition rule for mutually exclusive events})$$

In set theory notation, "OR" probabilities are referred to as "union" probabilities. The basic addition rule equation for two mutually exclusive composite events is written in set notation as:

$$P(A \cup B) = P(A) + P(B)$$

The addition rule can also be generalized to three mutually exclusive composite events as shown in the following set theory version of the equation

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

(*Equation 1.10a expanded for three composite events*)

Besides the additive composite probability of the three mutually exclusive joint events for "two heads in three tosses" just demonstrated, there is also an additive composite probability for three joint events that share the description of "one head in three tosses." The first of these joint events is the one for HTT, the fourth branch down in Figure 1.10. It is found to have a probability of 4/27 (Equation 1.10a).

$$P(\text{HTT}) = P(\text{H}) \times P(\text{T}) \times P(\text{T}) = \left(\frac{1}{3}\right) \times \left(\frac{2}{3}\right) \times \left(\frac{2}{3}\right) = \frac{4}{27}$$

The other two are the sixth and the seventh branches down from the top in Figure 1.10. They also each have a probability of 4/27.

$$P(\text{THT}) = P(\text{T}) \times P(\text{H}) \times P(\text{T}) = \left(\frac{2}{3}\right) \times \left(\frac{1}{3}\right) \times \left(\frac{2}{3}\right) = \frac{4}{27}$$

$$P(\text{TTH}) = P(\text{T}) \times P(\text{T}) \times P(\text{H}) = \left(\frac{2}{3}\right) \times \left(\frac{2}{3}\right) \times \left(\frac{1}{3}\right) = \frac{4}{27}$$

---

27 Mutually exclusive means that if one occurs the other cannot. Heads and tails are mutually exclusive. If a coin toss is "heads" it cannot also be "tails." If a three-toss joint event yields "3 heads" it cannot also yield "2 heads," since the "OR" is "exclusive or," exactly 2 or exactly 3. For our purposes, we will only consider mutually exclusive composite events. The equation for the general case addition ruleis: $P(A\_\text{OR}\_B) = P(A) + P(B) - P(AandB)$. here $P(AandB)$ is the probability of an event that qualifies as both event A and also as event B. Dropping this subtractive third term of the general case addition rule formula yields the formula for mutually exclusive events.

Consider a small sample of five digits "1, 2, 3, 4, 5." Define event A as "a digit greater than 2" and event B as "an even digit." With only these five digits to choose to draw from the probability of event A is $P(A) = 3/5$, since three of the five digits are greater than 2. The probability of event B is $P(B) = 2/5$ (since 2 and 4 are both even digits). One of the digits, 4, is both an even digit (event A) and also a digit greater than 2 (event B), so it might be counted twice. The general addition rule corrects for this: $P(A\_\text{OR}\_B) = P(A) + P(B) - P(AandB) = 3/5 + 2/5 - 1/5 = 4/5$. The special case formula for "mutually exclusive events" fails to correct for it: $P(A\_\text{OR}\_B) = P(A) + P(B) = 3/5 + 2/5 = 5/5$. It counts the 4 in both event A and event B without subtracting out the double count. The digit 1 doesn't fit either event, but the other four digits fit A only (3, 5) or B only (2), or both (4).

28 They are composite events in that each has a unique composition, a unique pattern of outcomes.

The additive composite probability of the three independent joint events for "one head in three tosses" is found also using Equation 1.10a.

$$P(1 \text{ head}) = P(A \cup B \cup C) = P(\text{HTT}) + P(\text{THT}) + P(\text{TTH}) = \frac{4}{27} + \frac{4}{27} + \frac{4}{27} = 12/27$$

The two single outcome branches of the eight in the branching diagram of Figure 1.10 are the first branch which has three heads in three tosses, and the eighth branch which has zero heads in three tosses.

$$P(\text{HHH}) = P(\text{H}) \times P(\text{H}) \times P(\text{H}) = \left(\frac{1}{3}\right) \times \left(\frac{1}{3}\right) \times \left(\frac{1}{3}\right) = \frac{1}{27}$$

$$P(\text{TTT}) = P(\text{T}) \times P(\text{T}) \times P(\text{T}) = \left(\frac{2}{3}\right) \times \left(\frac{2}{3}\right) \times \left(\frac{2}{3}\right) = \frac{8}{27}$$

We can now create an additive composite of all eight branches of the branching diagram of Figure 1.10, and we see that it is indeed a probability model with the sum of the probabilities being equal to 1.00.

$$P(\text{All 8 branches}) = P(\text{HHH}) + P(2 \text{ heads}) + P(1 \text{ head}) + P(\text{TTT})$$
$$= \frac{1}{27} + \frac{6}{27} + \frac{12}{27} + \frac{8}{27} = 27/27$$

We have just created a stochastic model of what is referred to as the *binomial distribution*.

1.3.6.3 The Binomial Distribution as a Simple and Clear Example of a Sampling Distribution of Proportions

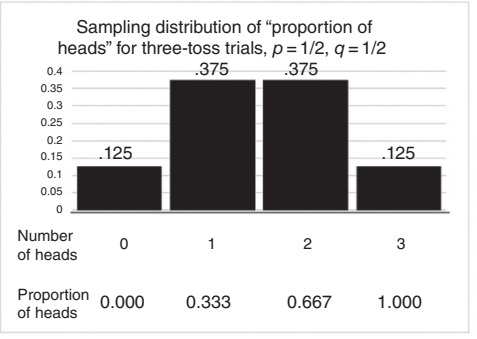
From the probabilities of the eight possible *independent joint events* in Figures 1.9 and 1.10, we can now create a sampling distribution of proportions for each of these two stochastic models. Although there are eight possible joint events for each, that is eight "branches" or patterns of results, there are only four possible outcomes in terms of the number of heads from these eight patterns for each model. The four outcomes are three heads, two heads, one head, and zero heads. The corresponding "mean number of heads" (that is, proportions) are 1.000 for the three heads outcome (HHH), $p = 0.667$ for the two heads outcome (HHT, THH, or THH), $p = 0.333$ for the one head outcome (HTT, THT, or TTH), and $p = 0.000$ for the no heads outcome (TTT).

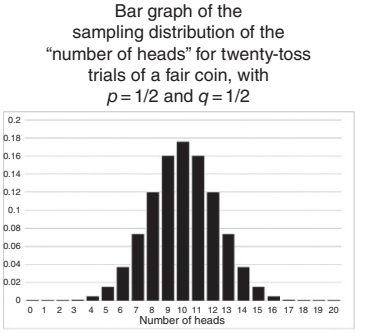
The probability model for each of these two stochastic models, one for the fair coin data and the other for the weighted coin data, can now be completed by creating a summary *sampling distribution of proportions* for each as shown in the two left panels of Figure 1.11. These two sampling distributions are an example of, and nothing less than an SDM, where the means are proportions, that is, the means of binary data. Each is a summary distribution of all the means that are possible for a given sample size, $N = 3$ tosses in this case, in sampling from a particular population. Taken together these identify the relative probability of each outcome.

It is unfortunate that the four "proportions" (the mean-number-of-heads values for each model) and the "probabilities" (the stochastic likelihood of the outcome of each) both start with the letter "*p*," making things a little confusing. The two panels, upper and lower, on the left side of Figure 1.11 provide a tabular comparison of the outcome probabilities of the fair coin model and the weighted coin model. The two bar graph panels in the middle of Figure 1.11, the fair coin one above and the weighted coin one below, provide a comparison of the shape of these two sampling distributions of the means of binary variables. The graphical distribution for the $N = 3$ fair coin model is obviously symmetrical but the one for the weighted coin model is clearly right-skewed. A small sample size of $N = 3$ was used to make the calculations simple and clear. But now, for a good graphical comparison

Sampling distribution of the "proportion of heads" for three-toss trials of a fair coin, with $p = 1/2$ and $q = 1/2$

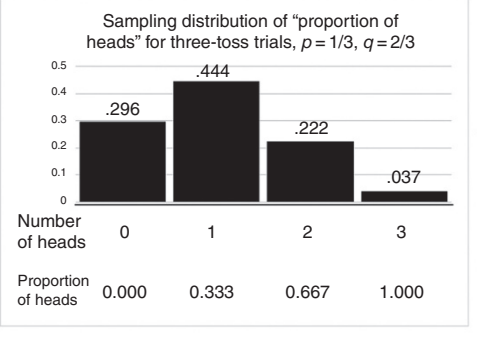
Number of heads	Proportion of heads	Probabilities (p)
3	1.000	p = 1/8 = .125
2	0.667	p = 3/8 = .375
1	0.333	p = 3/8 = .375
0	0.000	p = 1/8 = .125
		Sum = 1.000





Sampling distribution of the "proportion of heads" for three-toss trials of a weighted coin, with $p = 1/3$ and $q = 2/3$

Number of heads	Proportion of heads	Probabilities (p)
3	1.000	p = 1/27 = .037
2	0.667	p = 6/27 = .222
1	0.333	p = 12/27 = .444
0	0.000	p = 8/27 = .296
		Sum = 1.000



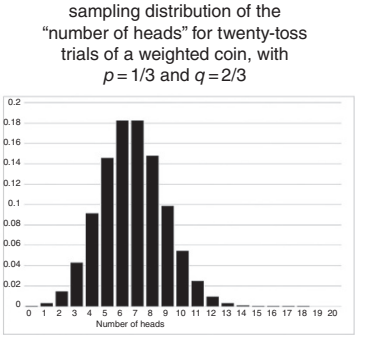


Figure 1.11 Summary tables and figures displaying the pattern of the "fair coin" results from Figure 1.9 appear in the upper three panels of Figure 1.11, with tables and figures of the "weighted coin" results from Figure 1.10 in the lower three panels. Tables of the sampling distribution of proportion of heads for the "fair coin" model compared to the "weighted coin" model are presented in the left panels of Figure 15, with a comparison of their respective bar graphs in the two middle panels of that figure. On the far right are bar graph representations for these two models applied to 20-toss trails, with the "fair coin" model on top and the skewed "weighted coin" model below.

of the sampling distributions, it would be good to have a larger sample. We therefore repeat these same calculations using a sample size of $N = 20$ both for the fair coin model and also for the weighted coin model, and obtain the resultant bar graphs shown in the two right panels of the figure. We expect the bar graph for the fair coin model to become more and more Gaussian as the sample size increases. It does. The $N = 20$ fair coin model graph in the top right panel of the figure is clearly bell-shaped. One could even say that, although it is a discrete distribution rather than a continuous curve, it approaches a proto-Gaussian shape. Interestingly, the corresponding $N = 20$ weighted coin graph in the bottom right panel of Figure 1.11 is also reasonably bell-shaped. The right skew of the probability model shows up primarily in the offset of the midpoint of the curve to the left side of the X-axis of the curve, centering on the $p = ⅓$ location in contrast to the $p = ½$ location for the fair coin graph.

The two-bar graph comparison on the right of Figure 1.11 is a good example of the *Central Limit Theorem*. It illustrates that regardless of the shape of the population distribution, the SDM from that population distribution will approach Gaussian shape as sample size N increases. The population for the weighted coin model is skewed with a 2/3 probability for tails, and a 1/3 probability for heads, but the sampling distribution of proportions is reasonably bell-shaped even with a sample size of only $N = 20$.

This section of Chapter 1 has a subtitle of "...a preliminary sidestep into statistical inference." Figure 1.12 continues that sidestep with a one-figure summary of the essential logic of statistical inference and sampling theory. Every statistical inference question involves data distributions of three kinds: (1) one or more samples, (2) the population distribution from which those samples are drawn, and (3) the sampling distribution of a particular summary statistic such as a mean or a proportion. Since a proportion is a special case of a mean, the process used for demonstrating statistical inference with proportions can also be applied to sampling distributions of means of fully quantitative data. Of these three types of distribution, the first one entitled "one or more samples" is real, that is, it is observed empirical data. The other two, the population and the SDM/proportions are hypothetical and mathematically constructed. Notice in this figure that example sample spaces are shown for three sizes of samples: a sample size of $N = 3$, a sample size of $N = 4$, and a sample size of 120. This figure summarizes the "big picture" of the logic of creating sampling distributions of proportions or of other types of means.

Examining the hypothetical sample space of the eight possible samples for an $N = 3$ binomial you will recognize it as a summary of what was shown in the branching diagram of Figure 1.10. The first of the eight icons in this sample space, the one at the top, has three heads on the right and zero tails on the left. Right below that icon is a set of three similar icons, identical to one another, each with two heads on the right and one tail on the left (but with each of the three representing one of the three different permuted orders of HHT, HTH, or THH[29]). Below that is another set of three icons. These have one head on the right and two tails on the left (HHT, THT, TTH). At the bottom is an icon with three tails on the left and zero heads. Below that sample space is its corresponding bar graph showing the shape of the SDM for this $N = 3$ binomial.

A similar sample space is shown to the right of that one, but for a sample size of $N = 4$. This sample space has sixteen possible sample outcomes. The first of the icons for this sample space, the one at the top, has four heads and no tails. Right below that is a set of four icons, each having three heads on the right and one tail on the left. Below this is six icons, each having two heads on the right and two tails on the left. Below this is another set of four icons with one head on the right and three tails on the left, and at the bottom is a single icon with four tails on the left and zero heads on the right. This $N = 4$ sample space also has its corresponding *SDM* summary bar graph below it. Each of the

29 As shown in Figure 1.10, these orders reflect the tosses (first, second, or third) on which heads or tails appear.

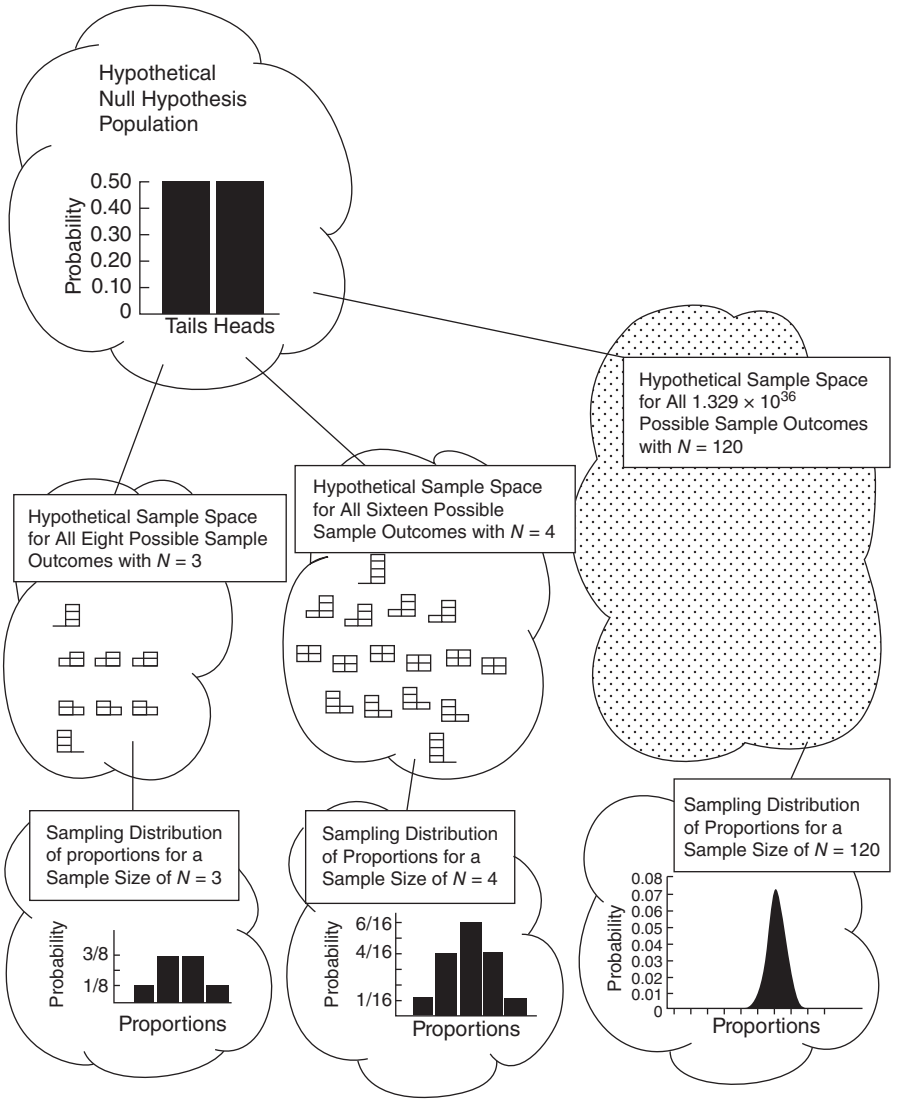


Figure 1.12 Single figure summary of the fundamental principle of statistical inference showing the three kinds of distributions: (1) a population, (2) samples from that population, and (3) sampling distributions of proportions from the sample spaces. *Source:* Reproduced from page 55 of Brown, Hendrix, Hedges, and Smith (2012).

bars in these bar graphs of the sampling distribution of means shows the relative frequencies (probabilities) of each binary mean (proportion) value.

1.3.6.4 Elaboration of Higher Level Calculations for the Binomial Distribution

The process of "expanding the binomial" was pursued early in the history of mathematics. To expand an expression such as $(p + q)^n$ means simply to multiply it out. Students in an elementary algebra class, for example, may learn to expand $(p + q)^2$ as $(p + q)^2 = p^2 + 2pq + q^2$. One can first multiply the $(p + q)$ binomial by p to obtain $(p^2 + pq)$. Next, multiply $(p + q)$ by q to obtain $(pq + q^2)$. Finally, sum the two conformable terms to obtain $p^2 + pq + pq + q^2 = p^2 + 2pq + q^2$.

This same process can be followed to obtain $(p + q)^3$ by multiplying $p^2 + 2pq + q^2$, first by p and then by q, and finally combining the two pairs of conformable terms to obtain the following equation.

$$(p + q)^3 = p^3 + 2p^2q + 2pq^2 + q^3$$

An astute French mathematician, Blaise Pascal (1623–1662)[30] noticed a pattern in these binomial expansions for various powers of $p + q$. When the expressions for $p + q$ taken to the first, second, and third power are placed above one another in a vertical array, with all the multiplicative coefficients showing, the pattern can be seen.

$(p + q)$ to the first power is $1p + 1q$

$(p + q)^2$ to the second power is $1p^2 + 2pq + 1q^2$

$(p + q)^3$ to the third power is $1p^3 + 3p^2q + 3pq^2 + 1q^3$

The coefficients for the algebraic expansion of the binomial for $(p + q)^4$ are shown in the Pascal's triangle of Figure 1.13 (which will be explained in the following paragraphs). As can be seen from the expansion of the binomial for $(p + q)^3$ and for $(p + q)^2$ all that is needed for the expansion of the binomial besides the coefficients is the appropriate powers of p and q as multipliers. Examining the "three-toss" binomial for $N = 3$ shown earlier, it is apparent that the appropriate powers of p start at N and decrease by 1 at each step from left to right, that is, from p taken to the third power (p^3) down

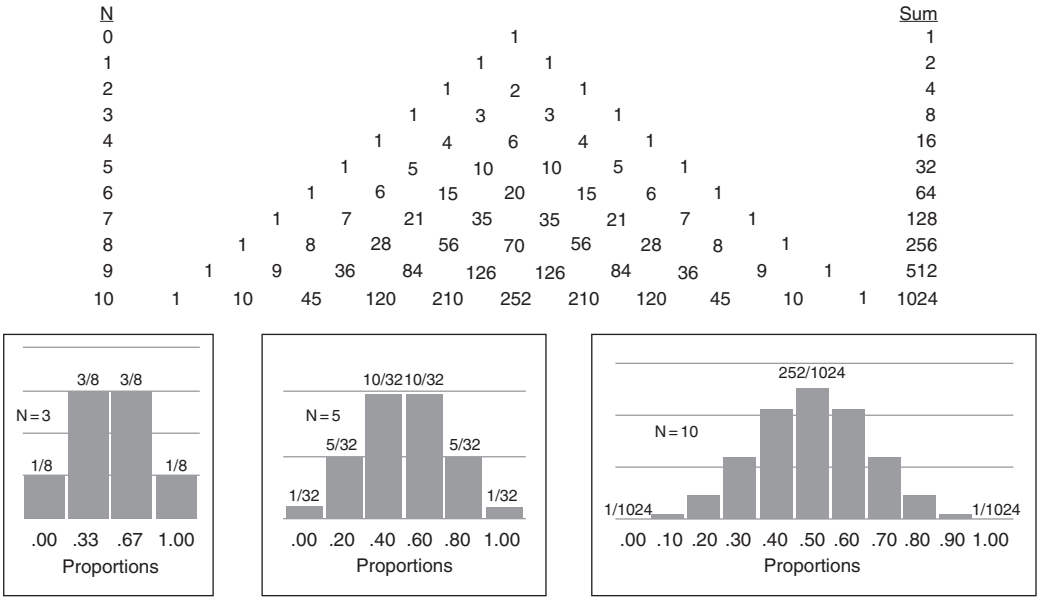


Figure 1.13 Pascal's triangle coefficients for all coin tosses beginning with $N = 1$ coin and extending to $N = 10$ coins. Also, three sampling distributions of proportions were created from these coefficients for samples of size $N = 3$, $N = 5$, and $N = 10$, with each having $p = 1/2$.

30 This had been noticed by several mathematicians in earlier centuries, but Pascal discovered it independently. In the western world we refer to it as Pascal's Triangle, but in China it is called Yang Hui's Triangle.

to p taken to the power of zero ($p^0 = 1$). The appropriate powers of q are seen to increase from left to right, going from $q^0 = 1$ up to q^3. Using that same procedure with the coefficients for $N = 4$ tosses, the binomial expansion looks like this:

$$(p + q)^4 = 1p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + 1q^4$$

When the appropriate values of p and q are inserted, this becomes the full mathematical expansion of the binomial. Suppose we are creating the binomial model for tossing a fair coin, with $p = ½$ and $q = ½$. The expansion of the binomial then returns the probability of each possible number of heads, from 0 to 4, that could be obtained in four tosses of the coin.

$$\begin{aligned}(p + q)^4 &= 1p^4 + 4p^3q + 6p^2q^2 + 6pq^3 + 1q^4 \\ &= \left(\frac{1}{2} + \frac{1}{2}\right)^4 = 1\left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right) + 6\left(\frac{1}{2}\right)^2 q\left(\frac{1}{2}\right)^2 + 4\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 + 1\left(\frac{1}{2}\right)^4 \\ &= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16} = \frac{16}{16} = 1\end{aligned}$$

These outcomes of 0, 1, 2, 3, or 4 heads are mutually exclusive and exhaustive and their probabilities sum to one, so this binomial expansion constitutes a very simple probability model.

Figure 1.13 can be used to explain the construction of Pascal's triangle, that is, the triangle of the multiplicative coefficients that are used to create the binomial expansion at each power of $(p + q)$. The first row of the triangle has the single value "1." The next three rows are the coefficients of the first, second, and third powers of $p + q$ just calculated. To the left of the Pascal's triangle is shown a column labeled *N*. This indicates the power of $(p + q)$ for each row of the binomial coefficients. On the right side of the triangle is shown a column of the sums (*Sum*) of the coefficient values for each binomial power *N*. Notice that the sum for each row is the number 2 raised to the *N*th power. For example, the row of binomial coefficients corresponding to $N = 3$ (which reads "1 3 3 1") sums to 8 which is equal to 2 raised to the third power, 2^3. To create additional rows, the value of each coefficient is found by simply summing the two coefficients on the row above it: the one to the left above it plus the one to the right above it. For example, the third coefficient in the row corresponding to $N = 5$ is 10, which is the sum of the two coefficients 4 and 6 directly above it, with 4 on the left above and 6 on the right above. This particular Pascal's triangle has been constructed from an *N* of one toss up to an *N* of 10 tosses. Each of these rows of the triangle can be used to solve a complex probability problem as shown in what follows.

For an illustration of a simple applied probability model, suppose you are a student taking a psychometrics course. When you arrive at an early morning class you are confronted with a six-item multiple-choice pop quiz to test your knowledge of yesterday's readings. There are four alternatives (a, b, c, or d) to choose from on each of the six multiple-choice items. Unfortunately, you neglected to do the readings yesterday, so you will have to depend on the "luck of the draw" to pass the quiz. What is the probability of your getting a score of 66.7% or better using guesswork on this six-item quiz? Given that there are four alternatives on each item, the population probability p of getting a correct guess on any one of them is equal to ¼ under the null hypothesis that you do not know the correct answers. The probability q, of *not* getting the answer correct is $q = 1 - p = ¾$. Since there are $N = 6$ items, you select the row of binomial coefficients corresponding to $N = 6$ (which reads "1 6 15 20 15 6 1").

There are seven possible scores that you could obtain on this six-item quiz: 0, 1, 2, 3, 4, 5, or 6 items correct. To obtain the probability values for each of these possible scores, multiply each coefficient

by the population parameters of p and q raised to their appropriate power. Examining the "three-toss" binomial for $N = 3$ shown earlier, it was observed that the appropriate powers of p start at N and decrease by 1 for each term from left to right, from p^3 down to p taken to the power zero ($p^0 = 1$). The appropriate powers of q, on the other hand, are seen to increase from left to right, from q^0 to q^3. Applying that same procedure to the $N = 6$ coefficients, the binomial expansion algebraically looks like this:

$$(p + q)^6 = p^6 + 6p^5q + 15p^4q^2 + 20p^3q^3 + 15p^2q^4 + 6pq^5 + q^6$$

Entering the p and q values into the equation will return the probability values for each of the seven possible proportions (means of binary data) of correct answers. The highest possible proportion of correct answers is all six out of the six items, which corresponds to a proportion correct of 1.00. The lowest possible proportion is none of the six items answered correctly, which corresponds to a proportion correct of 0.00. Applying the algebraic formula to the seven named proportions correct, from "P(6 correct)" to "P(none correct)" creates the following calculated probabilities (P) for each of the seven proportions.

$$\begin{aligned}(p + q)^6 &= P(6\text{ correct}) + P(5\text{ correct}) + P(4\text{ correct}) + P(3\text{ correct}) + P(2\text{ correct}) \\ &\quad + P(1\text{ correct}) + P(\text{none}) \\ &= \left(\frac{1}{4}\right)^6 + 6\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right) + 15\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^2 + 20\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^3 + 15\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^4 + 6\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^5 + \left(\frac{3}{4}\right)^6 \\ &= {}^{1}\!/_{4096} + {}^{18}\!/_{4096} + {}^{135}\!/_{4096} + {}^{540}\!/_{4096} + {}^{1215}\!/_{4096} + {}^{1458}\!/_{4096} + {}^{729}\!/_{4096} \\ &= {}^{4096}\!/_{4096} = 1.00\end{aligned}$$

The values of the seven binary means (the proportions correct) are mutually exclusive values, they are exhaustive of all the possibilities for six-toss trials, and the probabilities of the seven outcomes sum to 1.00. These are the three properties that constitute a probability model.

Our question was "what is the probability of getting a score of 66.7% or higher on the pop quiz using guesswork?" Four out of the six answers correct is a score of ⅔ correct, that is 66.7% correct, so the probability of getting a score of 66.7% or higher is the sum of the three probabilities, P(4 correct), P(5 correct), and P(6 correct). These three probabilities sum to ${}^{154}\!/_{4096} = {}^{1}\!/_{4096} + {}^{18}\!/_{4096} + {}^{135}\!/_{4096}$, which is 0.0376, or a little less than 4 in 100. Not a very likely outcome.

Section §13, the final section, will now close with the introduction of the multinomial distribution and its use in demonstrating the logical proof and the mathematical basis of two foundational principles of inferential statistics that were introduced at the beginning of this section. The first is the principle expressed in Equation 1.4, that the variance of means is equal to one Nth of the variance of raw scores, where N is the number of observations averaged. This variance of the SDM can also be referred to as the *squared standard error of the mean*. The second is the formula for calculating an unbiased estimate of population variance from sample data by using degrees of freedom in the denominator as shown in Equation 1.2:

$$\hat{\sigma}^2 = \sum x_i^2/(N-1).$$

1.3.6.5 Multinomial Demonstrations of Two Fundamental Principles of Statistical Inference

At the beginning of this final sub-section of the chapter, stochastic modeling is taken to a higher level by introducing the *multinomial distribution probability model*. This is similar in structure to the binomial distribution, but whereas the binomial is appropriate only for binary data, the multinomial distribution may be used with *discrete* quantitative data. The first of the two fundamental

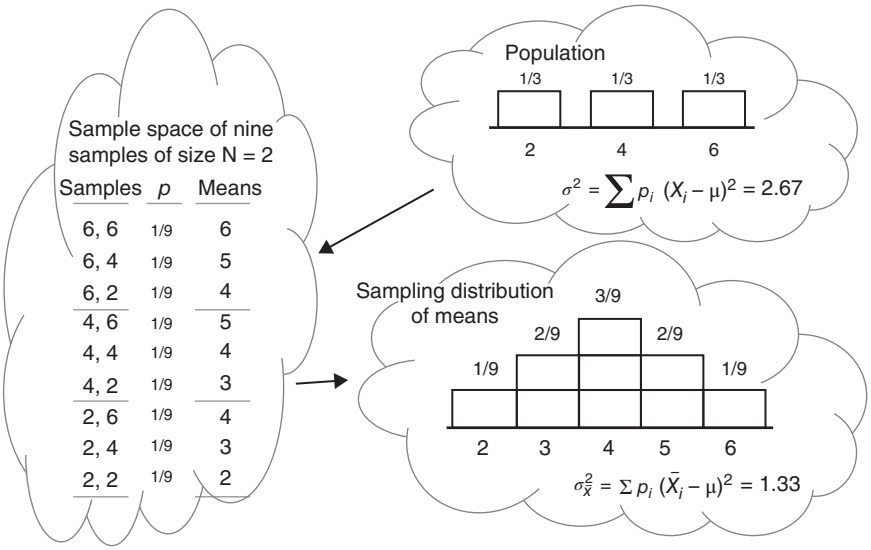


Figure 1.14 Diagram of the process of creating an SDM from a multinomial quantitative population distribution: (1) create a sample space of all the possible samples of size *N* from the population, with the probability and mean of each; (2) create a histogram displaying each of the means, labeled with its respective probability. (3) Calculate the variance of the SDM.

principles to be demonstrated is contained in Equation 1.4, the calculation of the variance of the SDM as one *N*th of the variance of raw scores, $\hat{\sigma}^2_{\bar{X}} = \hat{\sigma}^2/N$. This variance of the SDM, $\hat{\sigma}^2_{\bar{X}}$, is also the *squared standard error of the mean*, as stated in footnote 20 at the beginning of Section §13.

The first step in this demonstration is to create the simplest case multinomial population. This particular multinary population has only three values, 2, 4, and 6 as shown in Figure 1.14. Suppose these are pipes produced in a factory where there are equal numbers of pipes of each length produced. Each pipe length has a probability of $p = 1/3$ within the population. Since 2″, 4″, and 6″ pipes are equally likely in this population of pipes, the mean length is $\mu = 4$. The raw score variance is calculated using Equation 1.11. This is the "algebra of expectations" equation for a variance introduced in Appendix 2.

$$\sigma^2 = \sum p_i\,(X_i - \mu)^2 \qquad \textit{(Equation 1.11)}$$

$$\sigma^2 = \sum p_i\,(X_i - \mu)^2 = \left(\frac{1}{3}\right)(2-4)^2 + \left(\frac{1}{3}\right)(4-4)^2 + \left(\frac{1}{3}\right)(6-4)^2 = \frac{4}{3} + \frac{0}{3} + \frac{4}{3} = \frac{8}{3} = 2.67$$

The next step of the demonstration is to create the sample space of all possible samples of a given size from this population, including the probability (p) for each sample.[31] For the sake of simplicity of exposition, a sample size of $N = 2$ is used to create a sample space of only nine unique sample configurations, as shown on the left side of Figure 1.14. The third column in this sample space listing is the mean of each of the six possible means. There are five possible mean values, 2, 3, 4, 5 and 6, and their respective probabilities are 1/9, 2/9, 3/9, 2/9, and 1/9 as shown in the lower right of Figure 1.14. This is because each of the nine samples has a probability of 1/9, and there are three samples that have a mean of 4, two samples that have a mean of 3, two samples that have a mean

31 Since each sample has two pipes, such as 'two 6″ pipes in the first sample', 'a 6″ pipe and a 4″ pipe in the second sample', etc., and since each pipe length has a probability of 1/3, using the multiplication rule of "AND" probabilities, Equation 1.9, the probability value for each of the nine samples is 1/9.

of 5, one sample that has a mean of 6, and one that has a mean of 2. This creates the pyramid shape that can be seen for the SDM for these nine samples shown in the lower right of Figure 1.14.

The population has a mean of $\mu = 4$, and this is, of course, also the mean of the SDM $\mu_{\overline{x}} = 4$. Equation 1.12 can now be used to calculate the variance of the SDM, $\sigma^2_{\overline{X}}$.

$$\sigma^2_{\overline{X}} = \sum p_i \left(\overline{X}_i - \mu\right)^2 = \left(\frac{1}{9}\right)(2-4)^2 + \left(\frac{2}{9}\right)(3-4)^2 + \left(\frac{3}{9}\right)(4-4)^2 + \left(\frac{2}{9}\right)(5-4)^2 + \left(\frac{1}{9}\right)(6-4)^2$$

$$= \left(\frac{4}{9}\right) + \left(\frac{2}{9}\right) + \left(\frac{0}{9}\right) + \left(\frac{2}{9}\right) + \left(\frac{4}{9}\right) = \frac{12}{9} = 1.33 \qquad \textit{(Equation 1.12)}$$

The principle being demonstrated here is that the variance of the SDM is equal to one *N*th of the variance of raw scores in the population, where *N* is sample size, that is, the number of values being averaged to create the means. The variance of raw scores in this population of pipe lengths is $\sigma^2 = 2.67$, and the sample size for creating the SDM is $N = 2$. The variance of the SDM for a sample size of two should therefore be half of the variance of raw scores in the population as it is shown to be here: the population variance of raw scores is two and two-thirds and the variance of the SDM is exactly half that, one and one third. This same demonstration could be done with a sample size of $N = 3$ or of $N = 4$. Even a sample size of three would be quite tedious, with 27 unique samples in the sample space. A sample size of four would have 81 unique samples in the sample space. Rest assured that were this to be done, the variance of the SDM for a sample size of $N = 3$ would be $\hat{\sigma}^2_{\overline{X}} = \hat{\sigma}^2/N = 2.667/3 = 0.888$, the variance of the SDM for a sample size of $N = 3$ would be $\hat{\sigma}^2_{\overline{X}} = \hat{\sigma}^2/N = 2.667/4 = 0.666$, and if it were to be done for a sample size of $N = 10$, it would be $\hat{\sigma}^2_{\overline{X}} = \hat{\sigma}^2/N = 2.667/10 = 0.267$.

The second demonstration of a foundational principle in statistical inference using the multinomial will now be given. In the paragraphs leading up to Table 1.3 at the beginning of Section §13 two formulas for a variance were given. The first was Equation 1.1, $S^2 = \sum x_i^2/N = \left[\Sigma\left(X_i - \overline{X}\right)^2\right]/N$, where the sum of squared deviations from the mean was divided by *N*, the sample size to obtain an ordinary descriptive sample variance. The second formula, given in Equation 1.2, is the correct formula to use for statistical inference. This formula will provide an unbiased point estimate of the true population variance. This formula, $\hat{\sigma}^2 = \sum x_i^2/(N-1) = \left[\Sigma\left(X_i - \overline{X}\right)^2\right]/(N-1)$, uses degrees of freedom, $N - 1$, in the denominator in calculating variance rather than *N*.

It was there stated that Equation 1.1 is a biased estimator and that if it were used as the point estimate of the population variance, it would systematically underestimate the true population variance σ^2 by "one *N*th" of its value. That statement is exactly correct, as will now be demonstrated in Figure 1.15 and the accompanying explanation.

Figure 1.15 uses the same population of pipe lengths that was introduced with Figure 1.14, which has a population mean of $\mu = 4$ and a population variance of $\sigma^2 = 2.67$. The listing of the nine samples of size two, $N = 2$, shown in the leftmost column of Figure 1.15 will immediately be recognized as the same nine samples that were shown in the Figure 1.14 demonstration. Since they are the same nine samples, they will also each have the same sample means, and each of these nine samples has the same probability of $p = 1/9$ as those. The second column in Figure 1.15 gives the sum of squared deviations for each of the samples, that is, the numerator that will be used in calculating the variances. The first sample, "6, 6" obviously has a mean of 6 and the two scores have zero deviations from their mean. Therefore the sum of squared deviations is 0. The same is true of the fifth sample,

Listing of the nine N = 2 samples in the sample space	Sum of squares for each sample	Sample variance (S^2) from each sample	Unbiased estimate of population variance ($\hat{\sigma}^2$) from each sample
6, 6	0	0	0
6, 4	2	1	2
6, 2	8	4	8
4, 6	2	1	2
4, 4	0	0	0
4, 2	2	1	2
2, 6	8	4	8
2, 4	2	1	2
2, 2	0	0	0

Expected value of the sample variances across the nine samples

$$E(S^2) = \sum p_i S_i^2 = (3/9)(0)+(4/9)(1)+(2/9)(4) = 4/9 + 8/9 = 12/9 = 1.33$$

Expected value of the unbiased estimates of population variance across the nine samples

$$E(\hat{\sigma}^2) = \sum p_i \hat{\sigma}_i^2 = (3/9)(0)+(4/9)(2)+(2/9)(8) = 8/9 + 16/9 = 24/9 = 2.67$$

Figure 1.15 Multinomial demonstration comparing the expected value of the sampling distribution of S^2 (descriptive sample variances) with the expected value of the sampling distribution of $\hat{\sigma}^2$ (unbiased estimates of population variance).

"4, 4," and the ninth sample "2, 2." All three samples have zero deviations from their mean and therefore a sum of squared deviations of 0. Since this zero sum of squares happens three times out of the nine samples, the probability connected with the sum of squares value of $SS_1 = 0$ is $p_1 = 3/9$.

Examining the second column, it can be seen that four samples have the sum of squares values of $SS = 2$. These are sample 2 ("6, 4") and sample 4 ("4, 6"), both of which have a mean of 5 and therefore deviation scores of 1 and −1, and therefore a sum of squares of these deviations of $SS_2 = 2$. The same is true of sample 6 ("4, 2") and sample 8 ("2, 4"), except that the two of them have means of 3, but they also have deviation scores of 1 and −1, and therefore a sum of squares of these deviations of $SS_2 = 2$. Since all four of these have the same sum of squares equal to 2, the probability connected with this sum of squares value of 2 is $p_2 = 4/9$.

Again examining the second column, it can be seen that only two samples have the sum of squares values of $SS = 8$. These are sample 3 ("6, 2") and sample 7 ("2, 6"), both of which have a mean of 4 and therefore deviation scores of 2 and −2, and therefore a sum of squares of these deviations of $SS_3 = 4 + 4 = 8$, and the probability connected with this sum of squares value of 8 is $p_3 = 2/9$.

To summarize there are only three different sums of squares values among these nine samples, $SS_1 = 0$, $SS_2 = 2$, and $SS_3 = 8$, and their corresponding three probabilities are $p_1 = 3/9$, $p_2 = 4/9$, and $p_3 = 2/9$. The three descriptive sample variances are found by dividing these three sum of squares values by their N. The first one with a sum of squares of zero also has a descriptive sample variance of zero, $S_1^2 = \sum x_i^2/N = SS_1/N = 0/2 = 0$, and it happens three out of nine times, so its probability is $p_1 = 3/9$. The second one with a sum of squares of 2, has a descriptive sample variance of 1, $S_2^2 = SS_2/N = 2/2 = 1$, and it happens four out of nine times, so its probability is $p_2 = 4/9$. The third one with a sum of squares of 8, has a descriptive sample variance of 4, $S_3^2 = SS_3/N = 8/2 = 4$, and it happens two out of nine times, so its probability is $p_3 = 2/9$. The expected value, that is the mean, of these three descriptive sample variance outcomes should underestimate the true population variance by one Nth, which is by ½, since the N in each of these samples is two. That means it will be half as big as the true population variance. Since the true population variance is 2.67 (two and two-thirds), it should come out to 1.33 (one and one-third). Using the "expected values" formula for a mean from Appendix 2, that is exactly what it comes out to be:

$$E\left(\hat{S}^2\right) = \sum p_i S_i^2 = \left(\frac{3}{9}\right)(0) + \left(\frac{4}{9}\right)(1) + \left(\frac{2}{9}\right)(4) = \frac{4}{9} + \frac{8}{9} = \frac{12}{9} = 1.33$$

The unbiased estimates of the true population variance are found by dividing these three sum of squares values by $N-1$, their degrees of freedom. The first one with a sum of squares of zero also has an unbiased variance estimate of zero, $S_1^2 = \sum x_i^2/(N-1) = SS_1/(N-1) = 0/1 = 0,$ and it happens three out of nine times, so its probability is $p_1 = 3/9$. The second one with a sum of squares of 2, has an unbiased variance estimate of 2, $S_2^2 = SS_2/(N-1) = 2/1 = 2,$ and it happens four out of nine times, so its probability is $p_2 = 4/9$. The third one with a sum of squares of 8, has an unbiased variance estimate of 4, $S_3^2 = SS_3/N = 8/2 = 4,$ and it happens two out of nine times, so its probability is $p_3 = 2/9$. The mean, that is the expected value of these three unbiased variance estimate outcomes should agree with the true population variance of 2.67, which is exactly what it comes out to be:

$$E(\hat{\sigma}^2) = \sum p_i \hat{\sigma}_i^2 = \left(\frac{3}{9}\right)(0) + \left(\frac{4}{9}\right)(2) + \left(\frac{2}{9}\right)(8) = \frac{8}{9} + \frac{16}{9} = \frac{24}{9} = 2.67$$

Q. E. D., *quod erat demonstrandum*.

1.4 Summary

Chapter 1 begins with a discussion of why mathematical methods are central to the empirical sciences and particularly why statistical methods are used in the behavioral sciences and why we would use statistics rather than just a branch of applied mathematics. It then discusses the historical relationship between psychology and the emergence of statistical methods and how the Pearson product-moment correlation coefficient emerged from the work of Galton (perhaps the first psychometrician) and Pearson on the heritability of intelligence.

Section §6 then gives an account of the ongoing debate of the Ferguson Committee of the British Association for the Advancement of Science regarding whether the social and behavioral sciences could be considered to have bona fide scientific measurement, and a series of answers to that challenge. The first answer to the challenge came in the form of the influential 1946 paper by Stevens proposing four kinds of measurement scales, followed by a strong rebuttal to the Ferguson Committee challenge from Luce and Tukey (1964) using ACM, which in turn led to a series of three volumes on the foundations of measurement by Krantz, Luce, Suppes, and Tversky (1971, 1989, 1990). A brief account is also given of some encouraging and positive recent developments in the more general discipline of metrology and its most recent approach to measurement across the sciences.

The last part of the chapter deals with the development of scientific variables and the place of stochastic methods, comparisons of measures of central tendency and of dispersion in the development of psychometric methods. The chapter ends with a review of principles of probability and statistical inference that are foundational for psychometric work.

1.5 Study Questions

A. Discussion Questions

Q1 What is the foundational quantitative method for the past 120 years of development of psychometric theory and practice? Who was the author? *(the third paragraph of the introduction to Chapter 1)*

Q2 What multivariate analytical method was invented four years earlier in 1901 that led directly to the method described in question #1? Who was the author *(the third paragraph of the introduction to Chapter 1)*

§1 Why Are Mathematical Methods Central to the Empirical Sciences? Why Are Statistical Methods Used in the Behavioral Sciences? (p. 2)

Q3 Why have statistical methods become the primary quantitative approach within the behavioral sciences? *(all of Section **§1**)*

§2 The Specific Mission of Statistical Methods in the Behavioral Sciences (p. 3)

Q4 Explain the fundamental scientific functions of explanation and prediction. How do they differ and how does each contribute? *(Section **§2**, paragraphs 2 through 4)*

Q5 Compare the twin roles of statistical methods in scientific research, the exploratory role and the confirmatory role. How does science benefit from each? *(Section **§2**, paragraphs 2 and 3)*

§3 Why Do We Use *Behavioral Statistics* and Not Just a Branch of Applied Mathematics? (p. 4)

Q6 Explain the "historical development and mutual influence" of psychology and statistical science. *(all of Section **§3**)*

§4 Some Remarks on the Historical Relationships Between Psychology and Statistics: The Development of the Correlation Coefficient and Least Squares Regression (p. 5)

Q7 Who were Mendel and Galton? What is the difference between the Galtonian and the Mendelian mathematical approaches to genetics, as discussed by Sir R. A. Fisher (1918)? Describe how each of these approaches was carried out. *(Section **§4**, first paragraph)*

Q8 How was the *Pearson product-moment correlation coefficient* (PPMCC) developed? Who was its author? What were its historical antecedents? In what year was it published? *(Subsection **1.1.4.1**, paragraphs 1–2)*

Q9 When was the "least squares" mathematical method discovered, and by whom? Which very widely used contemporary statistical method grew out of it in the 20th and 21st centuries? *(Subsection **1.1.4.1**, third paragraph)*

Q10 Why is the symbol "*r*" used to represent the PPMCC? What does it stand for? *(Subsection **1.1.4.2**, fifth paragraph)*

Q11 To whom did Pearson initially attribute the conceptual/mathematical work that eventually led to his correlation coefficient? To whom did he finally give the credit in his 1920 *Biometrika* paper? Why did he change his mind? *(Subsection **1.1.4.2**, eighth paragraph)*

§5 On Terminology I: Attributes with Categories—Variables with Values—Tests with Scores. And What About "Data"? (p. 10)

Q12 Comment on each of the following three areas of semantic ambiguity identified by the authors in the terminology of behavioral research methods: the concept of "score," the

concept of "variable," and the concept of "data." *(first paragraph of Section **§5**; all of Subsection **1.2.1.1**, especially paragraphs 1, 4, 6, and 8)*

Q13 Summarize the three broad preconditions identified by the authors for producing bona fide research data. *(all of Subsection **1.2.1.2**)*

§6 Measurement Scales and the Representational Theory of Measurement (p. 13)

Q14 Explain how S. Smith Stevens' (1946) landmark paper on the four types of measurement, that is, the four measurement scales, grew out of the seven-year debate of the Ferguson Committee of the British Association for the Advancement of Science. *(all of Subsection **§6a** and Subsection **1.2.2.2**)*

Q15 What is meant by the term "ostensive fundamental measurement"? *(footnote 7 in Subsection **1.2.2.1**)*

Q16 Give data examples that fit each of Stevens' four types of measurement and identify the properties that distinguish among the four types. *(all of Subsection **1.2.2.3**, particularly the first five paragraphs and Figure 1.1)*

Q17 At which of the four measurement scales does Stevens say that "we come to a scale that is quantitative in the ordinary sense of the word?" On this basis, at what point should one place a "major dividing line" through the four scale types? *(Subsection **1.2.2.3**, the fourth paragraph)*

Q18 What is the central idea of the "relational theory of measurement" (RTM)? How did it evolve from its early beginnings in the late 19th century to the relational theory of Norman Campbell, to Stevens' popular redefinition as a theory of scale types, to its highly technical development in the axiomatic relational theory of measurement? *(Subsection **1.2.2.4**, particularly the first four paragraphs)*

Q19 What do Scott and Suppes (1958) and Domotor (1992) mean by "imbedding" and "injection"? What do they mean by projection? *(Subsection **1.2.2.4**, paragraphs 5 through 9).*

Q20 Compare the axiomatic RTM approach measurement to the huge academic industry of mainstream psychometric theory and practice including factor analysis, "true score theory," and CTT. How complex and conceptually challenging is each? How popular and practically useful is each? Where has each had its major impact? *(Subsection **1.2.2.5**)*

Q21 What is metrology and how has it affected the debate concerning the necessity of concatenation operations being essential to true scientific measurement? *(Subsections **1.2.2.6** and **1.2.2.4**)*

§7 *Scientific* and *Stochastic* Variables: When May "Data" Be Treated By Statistical and Test Theory Methods? (p. 29)

Q22 In what way is the analysis of test scores into the two components of "true score" and "measurement error," the stochastic component, central to CTT? *(Section **§7**, first paragraph)*

Q23 What is a random variable? What is the value of a truly random variable in quantifying the stochastic component of data? *(Section **§7**, paragraphs 2 and 3)*

Q24 In what way can a random process such as tossing two dice create a precisely specified distribution for the error component? *(Section **§7**, paragraphs 2 and 3, and Figure 1.6)*

Q25 Explain the *Central Limit Theorem* and how it relates to the sampling distribution of the "sums of the two dice values" shown on page 22. *(Section **§7**, paragraph 3, and Figure 1.6)*

Q26 What are the three necessary conditions for creating a systematic and precise probability model? *(Section **§7**, paragraphs 4 and 5, and Figure 1.6)*

§8 Group Level of an Attribute: The Location of the Group's Scores on an Interval Scale (p. 31)

Q27 How do histograms and means help in the description and comparison of two distributions? *(Section **§8**, paragraphs 1 to 3)*

Q28 Describe the two ways of calculating a mean, the traditional way of summing the scores and dividing by N, and the probabilistic way of multiplying each score by its "relative frequency," i.e. its probability of summing the products. *(Section **§8**, paragraphs 3 and 4)*

Q29 Describe the two ways of interpreting the fathers' mean head circumference (X) and the sons' mean head circumference (Y): the empirical/descriptive, and the formal/mathematical. *(Section **§8**, paragraphs 5 through 7)*

§9 Comparing median versus mean: Which one is "fair"? (p. 34)

Q30 What are the relative advantages and disadvantages of the mean and the median as measures of central tendency? *(Section **§9**, paragraphs 1 and 2)*

§10 A Psychometric Look at the Mean $\overline{X}$: Item *Difficulty* and p (p. 35)

Q31 Describe how the "oversimplified explanation of a performance score" is based on two concepts: the _____ of the tested subject, and the _____ of the task. How are these put together to create a dominance relation and to make inferences regarding the subject and the task? *(Section **§10**, paragraphs 1 and 2)*

Q32 How can a psychometric testing study be interpreted in two ways, as an agglomerative *attribute of a group of people*, or as *a property of the method, that is, of the test itself? (Section **§10**, paragraphs 2 through 6)*

Q33 How is a proportion calculated on a "dummy variable" (binary) related to the mean of a fully quantitative variable? *(Section **§10**, paragraphs 7 through 10)*

§11 Intra-group and Inter-individual Variability—*Dispersion* of Scores (p. 37)

Q34 Why are the standard deviation and variance the primary statistics for quantifying dispersion rather than the range? How are the range's characteristics like those of the median? *(Section **§11**, paragraph 4)*

Q35 Why are the standard deviation and variance the primary statistics for quantifying dispersion rather than MAD? *(Section **§11**, paragraphs 4 to 6)*

Q36 Why is the squaring of values a better way to deal with the problem of negative values rather than using absolute values? *(Section **§11**, paragraphs 6 and 7)*

§12 A Small Numerical Example, Finally (p. 40)

Q37 What is the approximate percent of the area under the normal curve found between plus and minus one standard deviation units from the mean? What is the approximate percent of the area between two and minus two standard deviation units? *(Section **§12**, paragraphs 3 and 4)*

§13 Does the Larger Head Circumference in a Sample of Sons than Fathers Hold for the Whole Population? A Preliminary Sidestep to Statistical Inference? (p. 40)

Q38 How is the unbiased point estimate of the population standard deviation calculated? How is it useful? Why is it so important to get the calculation right? *(Section **§13**, paragraphs 5 through 8, and Equation 1.3)*

Q39 How is the standard error of the mean calculated? How is it useful? *(Subsection **1.3.6.1**, paragraphs 1 through 4,and Equations 1.5 and 1.5a)*

Q40 What is a confidence interval for the point estimate of the population mean and how is it calculated? How is it useful? *(Subsection **1.3.6.2**, paragraphs 5 to 7, and Equation 1.6)*

Q41 How is the standard error of the proportion calculated? What is unusual about its calculation when compared to the standard error of the mean? *(Subsection **1.3.6.2**, paragraphs 1 through 5, Equation 1.8, and footnote 23)*

Q42 Explain the multiplication rule for independent joint events. How is it used in creating the binomial distribution? *(Subsection **1.3.6.2**, paragraphs 6 through 9, Equation 1.9 shown for both two and three joint events, Figures 1.9 and 1.10)*

Q43 Explain the addition rule for mutually exclusive composite events. How is it used in creating the binomial distribution? *(Subsection **1.3.6.2**, paragraphs 10 through 19, Equation 1.10 for two composite events and Equation 1.10a for three composite events, footnotes 27 and 28)*

Q44 How does the binomial distribution help to clarify the rationale and logic of an SDM? *(Subsection **1.3.6.3**, paragraphs 1 through 7, Figure 1.12)*

Q45 How is the SDM useful in statistical inference? *(Subsection **1.3.6.4**, paragraphs 1 through 8, Figure 1.13 and Figure 1.14)*

B. Exercise Questions

EQ1 A psychometrics class teacher gives a short eight-item multiple choice quiz at the beginning of class each morning to ensure students have read the material. The student must get a "pass" for each day by getting six or more of the eight items correct (75%) when choosing from among three alternative answers on each question. The binomial distribution can determine the probability of a student receiving a pass by random guessing alone. (The resources for building the binomial distribution for answering this question are found in Figure 1.3 and the accompanying text. Since there are 8 items in the quiz, go to the column labeled ***N***, then go to the row with an N of 8. Copy the nine coefficients to the right of it. The first of the nine is the coefficient for "0" items correct, the second for '1' items correct, the third for '2' items correct,...and the ninth for 'all 8 items correct.' These nine coefficients account for all nine of the possible outcomes in tossing a coin eight times or in our case guessing the answer to six or more items by chance. To construct the binomial distribution probabilities for each of the nine possible outcomes it is necessary to multiply *the* coefficient for each possible number of items 'guessed' correctly by its power of p and its power of q for each coefficient as outlined in Subsection ***1.3.6.3***, using the multiplication rule and the addition rule.)

EQ2 Use a spreadsheet to create a *Pascal's Triangle Calculator* following the instructions below. Use it to create the "bar graph of the sampling distribution of the proportion of heads for twenty-toss trials of a fair coin" ($p = 1/2$ and $q = 1/2$) as shown in the upper right panel of Figure 1.11 and compare your results with those in the figure."

Instructions for creating a *Pascal's Triangle Calculator* on a spreadsheet. This can be used for calculating binomial coefficients for binomial distributions with N values up to 20.

Use a spreadsheet such as Excel to create a Pascal's triangle calculator following the instructions below.

1. *Place a 1 in each cell of column A of a spreadsheet such as Excel, from cell A1 to cell A21.*
2. *Create a diagonal line of ones by going "down one and over right one" repeatedly from cell B2 to cell U21. The progression goes like this: B2, C3, D4, E5, ... down to ... Q17, R18, S19, T20, and U21. These two "lines" together with an underline across the bottom of row 21 will form a triangle that you can fill in.*
3. *Now begin to create the Pascal's triangle coefficients, beginning by entering the following into cell B3: "=A2 + B2". After hitting the "return" key, the number "2" should appear.*
4. *Copy the cell B3 you have just entered and paste its contents into row 4; cell B4 and then cell C4. A "3" should appear in each of them.*
5. *Now do the same thing to row 5, from B5 to D5. The numbers "4 6 4" should appear.*
6. *Continue doing this until you have completely filled the triangle.*

Now that you have successfully created your Pascal's Triangle Calculator, try it out by recreating the bar graph shown in the upper right corner of Figure 1.11 with a sample size of N = 20 and also with p = ½ and q = ½, a fair coin.

EQ3 *Now that you have created your Pascal's triangle calculator in Excel, and successfully used it to re-create the bar graph of the sampling distribution of the proportion of heads for 20-toss trials*

of a fair coin, shown in the upper right corner of Figure 1.11, do the same thing again, but this time with a weighted coin ($p = 1/4$ and $q = 3/4$) and compare to the lower right panel of Figure1.11. How does this sampling distribution of coin tossing with a p value of $p = 1/4$ compare to the one in the lower right corner of Figure 1.11 with a p value of $p = 1/3$?". Please use the same six-step instructions used for question EQ2 to now create this new Pascal's triangle for EQ3.

References

Alder, K. (2021). *The measure of all things: The seven-year odyssey and hidden error that transformed the world*. New York: Free Press.

Allen, M. J., & Yen, W. Y. (1979). *Introduction to measurement theory*. Long Grove, IL: Waveland Press, Inc.

Bravais, A. (1846). Analyse mathématique sur les probabilités des erreurs de situation d'un point. *Mémoires présentés par divers savants à l'Académie Royale des Sciences de l'Institut de France*, 9, 255–332.

BIPM (2019). (Bureau International des Poids et Mesures). *The international system of units (SI), 9th Edition*. St. Cloud, Paris, France: The International Bureau of Weights and Measures (BIPM). This SI Brochure is distributed under the terms of the Creative Commons Attribution 3.0 IGO https://creativecommons.org/licenses/by/3.0/igo/.

Brown, B. L., Hendrix, S. B., Hedges, D. W., & Smith, T. B. (2012). *Multivariate analysis for the biobehavioral and social sciences: A graphical approach*. Hoboken: John Wiley & Sons.

Brown, R. J. C. (2021). Measuring measurement – what is metrology and why does it matter? *Measurement (London)*, 168: 108408. https://doi.org/10.1016/j.measurement.2020.108408.

Campbell, N. R. (1920). *Physics: The elements*. Cambridge: Cambridge University Press. Reprinted (1957) as *Foundations of science: The Philosophy and theory of experiment*. New York: Dover.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London: Longmans.

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (2021). *The twenty-first mental measurements yearbook*. Lincoln: University of Nebraska Press.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, 3(3), 186–190.

Coombs, C. H. (1964). *A theory of data*. New York: John Wiley & Sons, Inc.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Daston, L. (1995). *Classical probability in the enlightenment*. Princeton, NJ: Princeton University Press.

Domotor, Z. (1992). Measurement from empiricist and realist points of view. In Savage, C. W. & Ehrlich, P. *Philosophical and foundational issues in measurement theory*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Ferreiro, L. D. (2011). *Measure of the Earth: The enlightenment expedition that reshaped our world*. New York: Basic Books.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52, 399–433.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. London: Macmillan & Co.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society*, 45, 135–145.

Galton, F. (1889). *Natural inheritance*. London: Macmillan.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley & Sons, Inc.

Harman, H. H. (1960). *Modern factor analysis*. Chicago: The University of Chicago Press.

Helmholtz, H. (1887/1971). An epistemological analysis of counting and measurement. In Kahl, R. (Ed.), *Selected writings of Hermann von Helmholtz*. Connecticut: Wesleyan University Press.

Hilbert, D. (1899/2015). *Grundlagen der geometrie*. Berlin: Springer Spektrum.

Hölder, O. (1901). Die axiome der quantitiit und die lehre vom mass. *Berichte uber die verhandlungen der Koniglichen Sachsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse*, 53, l–64.

Horst, P. (1965). *Factor analysis of data matrices*. New York: Holt-Rinehart & Winston.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2), 183–202.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, 2, 389–423.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2), 263–291

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Volume I: Additive and polynomial representations*. New York: Academic Press.

LaPlace, P. S. (1814/1952). A philosophical essay on probabilities. In F. W. Truscott, & F. L. Emory (Ed.), *Translated from the 6th French*. New York: Dover Publications.

Laplace, P. S. (1847/1995). *Théorie analytique des probabilités*. Paris: Editions J. Gabay, pp. 3–4.

Legendre, A. M. (1805). *Nouvelles methodes pour la determination des orbites des cometes*. Paris: Courcier.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Boston: Addison-Wesley.

Luce, R. D., Krantz, D. M., Suppes, P., & Tversky, A. (1990). *Foundations of measurement Volume III: Representation, axiomatization, and invariance*. New York: Academic Press.

Luce, R. D., & Narens, L. (1994). Fifteen problems concerning the representational theory of measurement. *Patrick Suppes: Scientific Philosopher*, 2, 219–249.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.

Majerova, V., & Majer, E. (1999). *Kvalitativni vyzkum v sociologii venkova a zemedelstvi. Cast I*. Praha, Czech: Ceska zemedelska univerzita. (Qualitative research in sociology of rural areas and agriculture. Part I.).

Michell, J. (1990). *An introduction to the logic of psychological measurement*. London: Routledge.

Michell, J. (2020). Thorndike's *Credo*: metaphysics in psychometrics. *Theory and Psychology*, 30(3), 309–328.

Mislevy, R. J., Almond, R. G., & Lucas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Report 632). Los Angeles, CA: CRESST.

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). Menlo Park, CA: SRI International.

Narens, L. (2002). *Theories of meaningfulness*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2010). The Rasch model and additive cojoint measurement. In M. L. Garner, G. Engelhard, Jr., W. P. Fisher, Jr., M. Wilson (Eds.), *Advances in Rasch measurement, Volume 1*. Maple Grove, MN: JAM Press.

Newell, D, & Tiesinga, E. (2019). *The international system of units (SI), 2019 edition*. US Department of Commerce, National Institute of Standards and Technology.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1), 1–18.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society*, 187-A, 253–318.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1), 25–45.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6* 2(11), 559–572.

Peirce, C. S. (1901/1957). The logic of abduction. In V. Tomas (Ed.), *Essays in the philosophy of science*. Indiana, IN: Bobbs-Merill.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3, 237–255.

Pfanzagl, J. (1959). *Die axiomatischen grundlagen einer allgemeinen theorie des messens. Schriftenreihe Statist. Inst. Univ. Wien, Vol. 1*. Würzburg: Physica-Verlag.

Pfanzagl, J. (1968). *Theory of measurement*. New York: Wiley.

Piovani, J. I. (2008). The historical construction of correlation as a conceptual and operative instrument for empirical research. *Quality & Quantity*, 42, 757–777.

Plackett, R. L. (1972). Studies in the history of probability and statistics. XXIX: The discovery of the method of least squares. *Biometrika*, 59(2), 239–251.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Savage, C. W., & Ehrlich, P. (1992). *Philosophical and foundational issues in measurement theory*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Scott, D., & Suppes, P. (1958). Foundational aspects of theories of measurement. *The Journal of Symbolic Logic*, 23(2), 113–128.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.

Spearman, C. (1904b). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15, 201–292.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 18, 161–169.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, 3, 271–395.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology*, 5, 417–426.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680.

Stevens, S. S., & Davis, H. (1938). *Hearing: Its psychology and physiology*. New York: Wiley.

Stewart, I. (2012). *In pursuit of the unknown: 17 equations that changed the world*. New York: Basic Books.

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 9, 465–474.

Stock, M., Davis, R., de Mirandés, E., & Milton, M. J. (2019). The revision of the SI—the result of three decades of progress in metrology. *Metrologia*, 56, 022001.

Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, E. Galanter (Eds.), *Handbook of mathematical psychology, Volume I*. New York: Wiley.

Suppes P., Krantz, D. H., Luce R. D., & Tversky, A (1989). *Foundations of measurement, Volume II; Geometrical, threshold, and probabilistic representations*. New York: Academic Press.

Thurstone, L. L. (1931). *The reliability and validity of tests: Derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems*. Ann Arbor: Edwards Brothers, Inc. (This is now unavailable.)

Thurstone, L. L. (1935). *The vectors of mind*. Chicago: The University of Chicago Press.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: The University of Chicago Press.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement*, 16, 8–13.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. *Science*, 185, 1124–1131.

2

Scores and Their Distributions

CHAPTER MENU

2.1 Quantification in Distributions: The Standard Deviation and Standard Scores, 65
2.1.1 (§14) A Psychometric Look at the Variance and SD: "Discriminability" of an Item, 66
2.1.1.1 Inappropriate Scoring, 67
2.1.1.2 The Standard Deviation and Binary Tests, 67
2.1.2 (§15) Using SD for Comparing the Incomparable: Percents vs. *Z*-Scores, 68
2.1.2.1 Norms and Baselines, 68
2.1.2.2 Using Percents with Norms and Baselines, 69
2.1.2.3 Comparing Scores Through Standardizing, 70
2.1.3 (§16) Density of Scores Along the Scale: The Normal Distribution of Frequency or Probability, the *N*-Law, 72
2.1.3.1 The Inverted Ogive, 72
2.1.3.2 The "Bell Curve", 73
2.1.4 (§17) The Bell Curve or Normal Curve Distribution Function $p(Z)$, 73
2.1.4.1 The Rationale Underlying the Simple Approximate Probability Distribution Equation, 77
2.1.4.2 The Need for a Cumulative Density Function, 77
2.1.5 (§18) The Ogive Curve or Normal Curve Cumulative Distribution Function $N(Z)$, 78
2.1.6 (§19) Percentile Position of a Score and the *N*-Law of Distribution, 80
2.1.7 (§20) Are *Z*-Scores and Percentiles Fair?, 81
2.2 Combining and Using Standard Scores, 82
2.2.1 (§21) More Convenient Scores Derived from *Z*: The Deviation *IQ*, *T*-Score, and Stanine, 83
2.2.2 (§22) What If the Scores Do Not Follow the *N*-Law? A Skewed Bell Curve, 84
2.2.2.1 Floor and Ceiling Effects: J-Shaped and L-Shaped Distributions, 86
2.2.2.2 Variables with Inherently Non-normal Distributions, 87
2.2.2.3 Population Selection and Extreme Groups, 87
2.2.3 (§23) Composite *Z*-Scores of Different Tests: Adding Apples and Oranges?, 88
2.2.4 (§24) Mutual Compensation, and Differential Importance of Sub-tests in the Composite, 90
2.3 Summary, 91
2.4 Study Questions, 93
References, 96

2.1 Quantification in Distributions: The Standard Deviation and Standard Scores

In the first section of the chapter, Section §14, the statistical standard deviation (SD) is examined as an attribute (the variability) of the group, and also as an attribute (the "discriminability") of the test. Test discriminability is the ability of the test to differentiate among individuals with respect to the

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

underlying property being measured. However, the SD represents a mixture of true individual differences and random or systematic error. Our senior author cautions against roughly categorizing continuous measures and encourages awareness of possible scoring-induced floor or ceiling effects. The SD of binary measures is a simple function of their mean or proportion. With binary data, test discriminability is maximized when the two alternatives are equally likely, and vanishes when either alternative always or never occurs.

The second section, Section §15, addresses the use of "percents" and standardized scores to provide comparability of different measures. It addresses common problems with the naïve application of percents to measures having different norms, baselines, and metrics. Cohen, Cohen, Aiken, & West (1999) proposed transforming measures to a scale they called the Percent Of Maximum Possible or POMP with 0 as the minimum possible value and 100 as the maximum possible value. This can be useful for many bounded measures, but unbounded measures are difficult to consistently express as percents without addressing the issues of norms and baselines. One common norm-based scaling approach is built upon the sample or population means and SDs. Standardized scores, also known as *Z*-scores, are formed by subtracting the mean from the observed value of each of the scores and then dividing these deviation scores by their SDs. Thus, *Z*-scores have a mean of zero and an SD of one.

The distribution of standardized scores is considered in Section §16, especially where these scores have a so-called normal distribution. In that case, more than 99% of scores will be found between ±3 SDs from the mean, capturing virtually the full range of the data. In Sections §16 to §19, the normal distribution probability density function (PDF) is described in terms of the familiar bell curve where the highest probability is at the mean, and smaller proportions are present for values farther below or above the mean. Alternate representations of the normal distribution in terms of its cumulative distribution function (CDF; or normal ogive), and the inverted CDF (or inverted normal ogive) are also presented and discussed. Finally, Section §20 considers whether percents, standardized scores, or normal percentiles provide a basis for "fair" comparisons among various measures. These alternative perspectives on the normal distribution can be described in familiar two-way function plots where the *X*- and *Y*-axes represent:

Function	Vertical *Y*-axis	Horizontal *X*-axis
Probability density function (PDF)	Probability density	*Z*-value
Cumulative distribution function (CDF) Or normal ogive	Cumulative probability	*Z*-value
Inverted normal ogive	*Z*-value	Cumulative probability

2.1.1 (§14) A Psychometric Look at the Variance and SD: "Discriminability" of an Item

In interpreting the variance or SD, there is an analogous situation to the two-way interpretation of the mean as a characteristic of *the sample* and of *the measuring instrument*. The established value of the variance, s_x^2 (or its square root, s_x, the SD), in a given group or population can be interpreted in two ways. On one hand, s_x^2 may be understood as a common *attribute of the group* of people, i.e. the group's variability, on the given test. On the other hand, it can be interpreted as an *attribute of the test*, one of its diagnostic qualities, for its use in a certain segment of the population (represented by the group in terms of a sample). Let us call the latter property the *discriminability* of the test; however, this is just a tentative term, without a standard use in psychometrics. Again, as in the case of

the mean, let's relate this psychometric interpretation of s_x^2 and s_x to the test's validity for diagnosing or measuring a targeted psychological attribute.

Even from an intuitive point of view, it is obvious that a test with almost *zero variance* cannot effectively *distinguish or discriminate* between people, for it assigns every tested individual almost the same score. By common sense, it cannot provide a sufficient basis for the validity of that test. On the other hand, we cannot make the opposite conclusion. A large SD does not guarantee the validity of a test. This is because a large dispersion of scores could be caused by a high level of unreliability of the test, by its high susceptibility to disrupting factors. The entire variance could thus be filled with error.

The content and scoring of a testing task should be manipulated by its developers to ensure sufficient variance in the population of interest, i.e. the test's "discriminability." This problem has been known from the early history of test theory. For instance, we may find the topic in the first year issue of the newly established *Psychometrika* journal in a paper by Marion Richardson (1936) who used the term "differential validity" (this term has also been recently used with a different meaning, see §28).

2.1.1.1 Inappropriate Scoring

Sometimes even a test with good conditions for having plausible psychometric properties may lose them through inappropriate scoring. One way this could happen is when the detailed original values are agglomerated into a small number of categories with the justification that they were "more practical" to use as test scores. For example, the original values, given as the number of produced words in the *V-test*, from 0 to 35 words, could be automatically converted by an administrator of the test into just a 4-category scale, and directly recorded only as numbers from 1 (for very bad) to 4 (for excellent). The following scheme might be suggested: the score up to 5 words would be coded by "1," from 6 to 15 by "2," from 16 to 30 by "3," and over 30 by "4."

Besides this distortion of the original metric of the scale, retaining only the ordinal character of the four categories and changing the frequency distribution, the more detailed information from the original values and knowledge about their frequencies would also be lost. And also of importance here, there would be a substantial decrease in the variance and SD as well.

2.1.1.2 The Standard Deviation and Binary Tests

Unlike mathematical quantities with unlimited possible values ranging from $-\infty$ to $+\infty$, behavioral variables are very often influenced by "floor" and "ceiling" effects by virtue of their scoring. Every *binary* (*dichotomous*, Yes/No) test is influenced by this effect which produces a certain fixed *relationship* between the *mean*, $\bar{y} = p$, as an index of test difficulty, and the *variance*, s_y^2, as an index of test "discriminability." This relationship is obvious in the cases where binary test scores are quasi-quantified as either $Y = 1$ or $Y = 0$. On the same principle by which the mean of such a test equals the proportion of "good" scores, i.e. $\overline{Y} = p$, it is also possible to verify that its variance can be written in a special simple form $s_y^2 = p(1-p)$. Using the notation $q = (1-p)$ we have $s_y^2 = pq$ and $s_y = \sqrt{pq}$. Thus, s_y^2 as an index of test "discriminability" with binary tests becomes an exact function of its difficulty, p.

You can easily verify that the binary test has its maximum variance, i.e. its maximum "discriminability" if its difficulty is $p = 0.5$. Then the maximum of $(p \times q) = 0.25$. Any difficulty other than $p = 0.5$ will yield a smaller variance value since the proportion of "good" and "failed" scores match the condition $p + q = 1$. For example, $p = 0.6$, and $q = 0.4$ give us $s_y^2 = 0.24$ etc. Maximum p is

reached if all people are scored 1, then $p = 1$, and variance, i.e. "discriminability," is $s_y^2 = 0$. Of course, s_y^2 also equals 0 if $p = 0$, that is if $q = 1$ when all people failed the test. In both cases, the test does not yield any information useful for distinguishing between people in the given group or population. Such a test is either absolutely difficult or absolutely easy for the given population category, therefore its zero variance means zero "discriminability" and, finally, zero validity—hence, the test is absolutely worthless as a diagnostic method in the given subpopulation.

2.1.2 (§15) Using SD for Comparing the Incomparable: Percents vs. *Z*-Scores

A specific class of problems for comparisons is related to the use of percents. Typical illustrations of these problems relate to using an interval scale with no objectively established zero—"natural zero" or another kind of "fixed zero"—on the scale. A usual didactic example can be shown with temperature scales. Say that yesterday morning I saw a 10 °C reading on my window thermometer with the Celsius scale. But if I read 15 °C this morning, it makes no sense to say "it is 50% warmer today."

Now consider percents, which have another kind of problem. Imagine two sports fans, A and B, arguing about which of their favorite athletes made better improvements from last year to this year. What is a larger increment of performance—that of a top sprinter, the favorite of A, whose result in the 100 m sprint was improved from 10.0 to 9.9 seconds or that of a beginning Marathon runner, whose advocate was B, who improved from 5 : 00 to 4 : 57 hours, i.e. from 300 to 297 minutes?

Although most would agree that the improvement of the sprinter is more remarkable than that of the Marathoner, the two fans still started to seek some quantitative measure that would permit a mathematical, and hopefully "objective," comparison. Although both performances were measured in the same physical quantity—expressed by time units—the two friends still felt these were not directly comparable. One approach might be to simply compare the increments, 0.1 seconds to 3 minutes, in terms of percents. Why not, percents are often assumed to be "unit-free." To the big surprise of A, who was sure that any mathematical and objective method must support the obvious excellence of the sprinter's improvement, they learned that the two increments are the same—each represents an improvement of 1%.

Nevertheless, we still feel that something is wrong here. What could make different values—even in different units of measurement—comparable? What could have possibly led the two fans to think about percents? They hoped to find a common base across the two different scales—in fact, two values that are different but based on the same logical background. They were looking for baselines comparable by common conceptual content. The appropriate common external bases that make different things comparable are sometimes called *norms*. Comparing things with respect to norms is a usual part of everyday life, including qualitative norms such as comparing someone's behavior to social or cultural norms. Quite often, the norms are represented by a just acceptable minimum or by a desirable or recommended limit, like the nominal Recommended Daily Intake (RDI) of nutrients in terms of percents.

2.1.2.1 Norms and Baselines

To stay with the women's marathon example, we might compare a theoretical improvement of a world record through one minute in two equal time periods from the past. During the six-year period from 1971–1977, the world records changed from 3 : 01 : 42 to 2 : 34 : 48, about 25 minutes. Later, from 1996–2002, a six-year period again, the performance grew closer to an ideal ceiling, hence the records moved from 2 : 20 : 17 to 2 : 17 : 18, i.e. a range of only three minutes.

However, one minute is proportionally very small to any of the two starting baselines, 182 minutes in 1971, and 140 minutes in 1996:

One-minute improvement in the former period represents 1/182 ... 0.6%,
One-minute improvement in the latter period represents 1/140 ... 0.7%.

Hence, a wrong conclusion could be made that a one-minute improvement was proportionally almost the same. Such an approach omits the fact that in the former period, the improvement of a record by one minute represented just 1/25, i.e. 4% of the range in which the records had improved. However, in the latter period, the one minute represented 1/3, i.e. 33% of the whole range. It means that it was "more difficult" to move the record by one minute in the years 1996–2002 than it was in 1971—and the percentage comparison presented earlier definitely does not represent this difference appropriately. A better baseline would be the range of scores for each of the periods because the range puts the increment in context.

2.1.2.2 Using Percents with Norms and Baselines

We can find such mistaken usage of percents in many other areas of data evaluation, including medicine, physiology, economics, etc. Some additional examples corresponding to the aforementioned types are briefly described as follows.

On the *absolute location of a baseline* with respect to the *variable-as-intended*, i.e. different than as it is measured: No newborn has zero weight, no living human being has zero body height, zero number of erythrocytes, zero blood pressure, etc. It is confusing to say that the intraindividual oscillation of *IQ* is ±3% since there is nobody alive with $IQ = 0$ as a baseline. Moreover, the contemporary *IQ* is already expressed relatively, namely with respect to a baseline $IQ = 100$ anchored in an average of the "normal population" with its approximate minimum at $IQ = 55$ (i.e. the average *IQ* of $100 - 3$ SDs, where each $SD = 15$, see later Section §22).

Completely failing in a didactic mathematical test with a score of 0 does not mean an absolutely zero amount of mathematical ability—perhaps the test was too difficult or maybe it was constructed for a different population. From the other end of the scale: Nobody has a reaction time of 0 second, which example shows a baseline located at the *maximum* performance instead of the *minimum*, although the problem remains similar. Another example of using an average as a baseline is *LD*, the Lethal Dose of a drug. It definitely should not be established based on a very extreme case of the only exceptional subject who has survived such a dosage—instead, it is evaluated according to a normal probability distribution in a similar way to *IQ*, i.e. as the mean. On the other hand, in medicine, some baselines are very often stated in the form of a *desired nominal value*, usually combining a recommendation of experts with normal values found in reality, such as the *RDI* of vitamins, or maximum tolerable dosage of radiation, etc.

On the *relative comparability of two baselines*: It is necessary to check whether two numerical values of percents have been established *on the same principle*, i.e. whether they were related to the same semantic and logical type of their corresponding baselines. One should not be related to a minimum (or maximum) and the other one to an average, and/or to a nominal value stated by experts—such percents would be incomparable.

On *range*-related aspects: None of the carefully treated aspects mentioned here can circumvent the problem of range. A certain percentage may be standard in a certain range of absolute values but cannot be applied outside of that range. In a conference on obesity, a nutritionist may state that it is generally accepted that a reduction in body weight by more than 5% is "not healthy." However, the expert was not able to explain the case when a man weighing over 400 kg (unable to get through the door, who was taken out through a window by a crane) and was moved to undergo treatment to

reduce his weight to 120 kg. The nominal *RDI* for vitamin C is 60 mg (milligrams), though without any problem a person might take a soluble pill containing 1000 mg exceeding about 17 times the recommended value (baseline), i.e. 1700%. Now compare this with the nominal *RDI* for selenium in your food, whose *RDI* is up to 200 μg (micrograms), but when exceeded only 5 times, i.e. 500%, selenium becomes strongly poisonous and would make you sick, lose your hair, etc., and nobody would attempt a daily intake of 17 times the *RDI* as they might with vitamin C. Hence, a percentage with respect to an absolute baseline is confusing if not accompanied by information about the range.

The appropriate common external bases that make different things comparable are often called norms. Certainly, comparing things to norms is a usual part of everyday life, including not just quantitative norms but also *qualitative norms* such as comparing someone's behavior to social or cultural norms. Quite often, these are established as *desirable norms* or a recommended limit, like the RDI, speaking of a *norm as something which should be*.

2.1.2.3 Comparing Scores Through Standardizing

Having established a norm using the statistical mean, norm-related comparisons of two different things are easily done if one value is "above its mean" and the other "below its mean." However, what if both values are respectively either "above," or "below?" How can we decide which value is "more distant" from the mean? To answer this question, let us compare scores in two subtests that are parts of the battery for neuropsychological assessment of Alzheimer's Disease—AD, as developed by the Consortium to Establish a Registry for AD (CERAD). Let's use the reference values from a large normative study by Welsh et al. (1994), with means and SDs for different age and sex categories. One subtest, called *Constructional Praxis*—or just *C-Praxis*—is intended to measure a visuospatial-constructional ability and its score starts at 0. This is derived from the number of objects a tested individual is able to express by drawing. The other subtest is *Verbal Fluency,* or *V-Fluency*, for short. This subtest was created to measure language and verbal productive ability, its score is expressed as the number of words (names of animals) a tested individual is able to name in one minute. Now, imagine we are in the position to compare the two tests' scores x and y:

i) achieved by an 80-year-old woman in *C-Praxis* as X, namely $X_w = 9$ [drawings],
ii) performed by a 50-year-old man in *V-Fluency* as Y, namely $Y_m = 23$ [words].

Until we know the "norms," such a task would be nonsense. However, knowing that CERAD data from a survey of several hundred people in each category is used by clinicians as a reference standard (cf. Welsh et al., 1994), with arithmetic means for the corresponding genders and ages where $\overline{X} = 10$ and $\overline{Y} = 18$, we easily decide that $X_w = 9$ is worse than $Y_m = 23$. Of course, this is not because $9 < 23$, but because $X_w = 9$ is below its category mean, while $Y_m = 23$ exceeds the corresponding category mean.

Geometrically, we have transformed the two scales on the axis of real numbers by shifting or *transforming* them such that their *origins*, or *zero points*, have been *centered* at their corresponding means. Therefore, the original values from the zero point are now expressed as the *deviations* from their means, and the new test scores are now called *deviation scores* or *centered scores*, designated by $x_i = \left(X_i - \overline{X}\right)$. Thus, we can immediately see if the centered deviation scores are either positive or negative, i.e. above or below the mean.

Now, what if both values were "on the same side" of their "norms?" For instance, if we were to compare the values

$$X_w = 9\,[\text{drawings}] \;\text{ to }\; Y_m = 12\,[\text{words}],$$

or in deviations:

$$\left(X_w - \overline{X}\right) = -1\,[\text{drawing}] \quad \text{to} \quad \left(Y_m - \overline{Y}\right) = -6\,[\text{words}],$$

i.e. when both values are below their corresponding means. Then we need to decide which of them is "farther below its mean" than the other. It means that the transformation of the scale origins to the means was not enough to make the decision. We also need to introduce a *transformation of the units* of measurement for each of the two scales to make them comparable. That is, we would like to get units that would be comparable and would reflect the different amounts of variability in the two scales. Somehow, the new units should exhibit the position of a score between its mean and its minimum (or maximum). This would show which score is farther below (or above) its mean: the *new* comparable *units of measurement* are *standard deviations*, *SD*s. Therefore, we would need more complete information from the standardization study, including the SDs, i.e.

$$\begin{array}{ll} \overline{X} = 10\,[\text{drawings}] & \overline{Y} = 18\,[\text{words}] \\ s_x = 1\,[\text{drawing}] & s_y = 5\,[\text{words}] \end{array}.$$

(These are actual values from a standardization study; however, they have been rounded.)

Now we can determine the number of SDs that lie between the original raw score values and their corresponding means. These new "normalized scores" are sometimes called "sigma-scores," but more typically *Z-standardized scores* or more shortly *Z-scores:*

$$\begin{array}{ll} z_x = \left(X_i - \overline{X}\right)/s_x & z_y = \left(Y_i - \overline{Y}\right)/s_y \\ \quad = (9-10)/1 & \quad = (12-18)/5 \\ \quad = -1 & \quad = -1.2 \end{array}$$

Because the younger man's score $z_y = -1.2$ in *V-Fluency* lays 1.2 SD below his category mean, it represents "lower performance" than the older woman's score $z_x = -1.0$ in *C-Praxis*; her *intragroup relative position* is just 1.0 SD below her category mean.

Notice also that if we were speaking the language of measurement of physical quantities, the *Z*-scores could be recognized as "quantities without dimension," much like percents. The deviation −6 [words] divided by an SD of 5 [words] yields the value −1.2, which seems not to have any real-world unit of measurement. Therefore, comparing the *Z*-scores of scales of different real meaning or content does not represent "counting apples and oranges together." Here you just count together—or mutually compare—things that do not have any real names, and this fact makes them comparable, and countable together. This is very important for psychological assessment. Test batteries such as the CERAD consist of tests or items where each is measured in different units, say words, drawings, points, seconds, inches.... Without having the "unitless" *Z*-scores, how could you create and compute a *composite total score* by adding "apples and oranges," the scores of different tests?

By tradition, the original scores *X*, *Y*, ... obtained directly by diagnostic methods such as tests, inventories, and measurements are called the *raw scores*. They are expressed in measurement units such as seconds, inches, etc., or in the number of correctly answered questions, items, or the number of mistakes, etc. On the other hand, besides *Z*-scores, there are other kinds of transformations of the raw scores, which may yield some other directly comparable values. Their transformations into *Z*-scores or other kinds of comparable values are then known as different types of *standardized scores*. Some additional types of transformations other than *Z*-scores will be introduced in Section §21.

2.1.3 (§16) Density of Scores Along the Scale: The Normal Distribution of Frequency or Probability, the *N*-Law

Although a comparison based on *Z*-scores might seem to be related only to the mean as a norm, having little to do with the score's *position between its minimum and maximum*, this is not true. In most cases, the observations or measurements of many attributes in psychology show a certain lawfulness, or we may say they are governed by a statistical law, the well-known ***N**-law*, the law of "normality." A consequence of the ***N***-law tells us that the minimum value in a data set is approximately located 3 SDs below the mean, while the maximum is located approximately 3 SDs above the mean. Therefore, the minimum is approximately assigned $-3z$ points while the maximum is assigned approximately $+3z$ points on the *normalized* scale of *Z*-scores. According to the statistical theory of the ***N**-law*, this means that approximately 99% of the data values lie between −3 and +3 SD points from the mean, that is, almost 100% of them.

Also, what do we call the interval pictured in Figure 2.1 and containing all the data values? Yes, it is the *range:* $\mathrm{Ra}(x) = x_{max} - x_{min}$. Therefore, the statistical relationship between the range Ra and the standard deviation SD, namely that $\mathrm{Ra}(X) \approx 6s_x$, or that $X_{max} = \overline{X} + 3s_x$ and $x_{min} = \overline{X} - 3s_x$ is called the *rule of* ±3 *SD* or the *rule of 6 SD*. Sometimes a kind of terminology which mixes the SD in the sample, s_x, and in the population, σ_x, is being used, and therefore the relationship has become perhaps more popularly referred to as *the rule of six sigmas*. Quite often, especially in smaller samples, we are satisfied with the use of only a *rule of* ±2 *SD* to approximate the minimum, maximum, and the range (see also Section §18). These properties of the SD represent another illustration of why the SD is so useful and is to be preferred over the MAD, the Mean Absolute Deviation.

2.1.3.1 The Inverted Ogive

In this section, Blahuš relates that as a student, he was fascinated when his professor in a theory of education class, who was not a statistician, introduced the class to a series of mutually very similar curves illustrating several profiles of the frequencies of the students' scores in tests of quite different

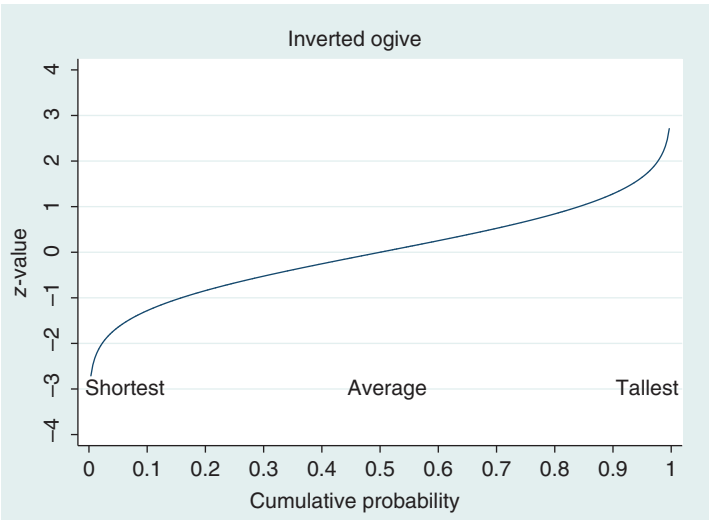


Figure 2.1 Empirical form of the inverted normal ogive.

nature. Where did the similarity come from? The answer was that this was due to the "law of normality," how randomness works in nature for most of the attributes you can measure.

For those who were still hesitating, a small experiment was immediately presented. The professor predicted the form of another curve that he claimed was an effect of the same law and it was to be confirmed by the "experiment." There were about 50 students in the lecture hall, and they were asked to stand in a line ordered according to body height from the smallest to the tallest along a wall. Then each stepped out for a second to watch the curve created by the tops of their heads along the wall. And, yes, the prediction has been confirmed, the tops of their heads projected on the wall created a lazy S-like horizontal curve as in Figure 2.1. The S-like form called "inverted ogive" (the term being imported from architecture where it describes certain Gothic arches) can be "normally"—meaning very often—found in nature. This inverted ogive is also known as the inverse normal cumulative distribution.

Dr. Blahuš also relates that when his children were small, they liked to arrange leaves fallen from a tree onto the ground to repeatedly find the S-curve for leaves ordered by size being interestingly the same for different species of trees. This is because the extreme cases—extremely small or extremely large—are rare, they are not frequent, maybe they are not quite "normal," and therefore those bend the ends of the curve in Figure 2.1 very quickly down or up. What is "normal" is frequent, typical, or "modal" in the sense of the definition of mode, and therefore the series of more frequent "normal" cases makes the middle part of the inverted ogive so long.

2.1.3.2 The "Bell Curve"

An empirically demonstrated inverted ogive is probably the easiest way to illustrate the occurrence of the *Law of Normal Distribution of Probability*, the ***N**-law* in nature, but there is another form by which the law is usually introduced in textbooks, namely the "bell curve." While the empirical inverted ogive expresses the properties of the sequentially accumulated frequencies, the bell curve (Figure 2.2a) represents a picture of a series of noncumulative frequencies in each interval along the scale. The proper name for it in its theoretical form is the *Gaussian* or Laplace-Gauss *normal pdf* (probability distance function) since with increasing sample size, *N*, the empirical discrete step frequencies are in theory supposed to converge to a smooth bell-shaped curve of probability as indicated in Figure 2.4.

Figure 2.3 compares the inverted normal ogive curve that typifies the horizontal S-curve (for leaves ordered by size and the heights of students in a class) with the CDF as shown on the lower half of Figure 2.2.

2.1.4 (§17) The Bell Curve or Normal Curve Distribution Function *p*(*Z*)

What is the famous mathematical function describing the Gaussian "bell curve"? In other words, how does one compute the *probability density*[1] function values that correspond to the value of each *X* data point? The fundamental formula applies to continuous variables (an exception for our usual approach of using integer data) and looks complex enough to be disregarded by every

1 When we create a "standard normal curve" like Figure 2.4, it is done by fitting a continuous curve to the vertical location corresponding to each *Z*-score location along the abscissa. These points are not probability values, but so-called *probability densities*. Probabilities refer to intervals, not to numerical locations on a continuous scale, such as the normal curve scale shown in Figure 2.4, where there are an infinite number of possible locations between any two points. When data are discrete, such as number of heads in seven tosses of a coin, there is an interval (between 6.5 and 7.5) and a probability can therefore be calculated. Asking what the probability is of getting a "*z* test value of 1.645 or above" specifies an interval and its probability can be identified.

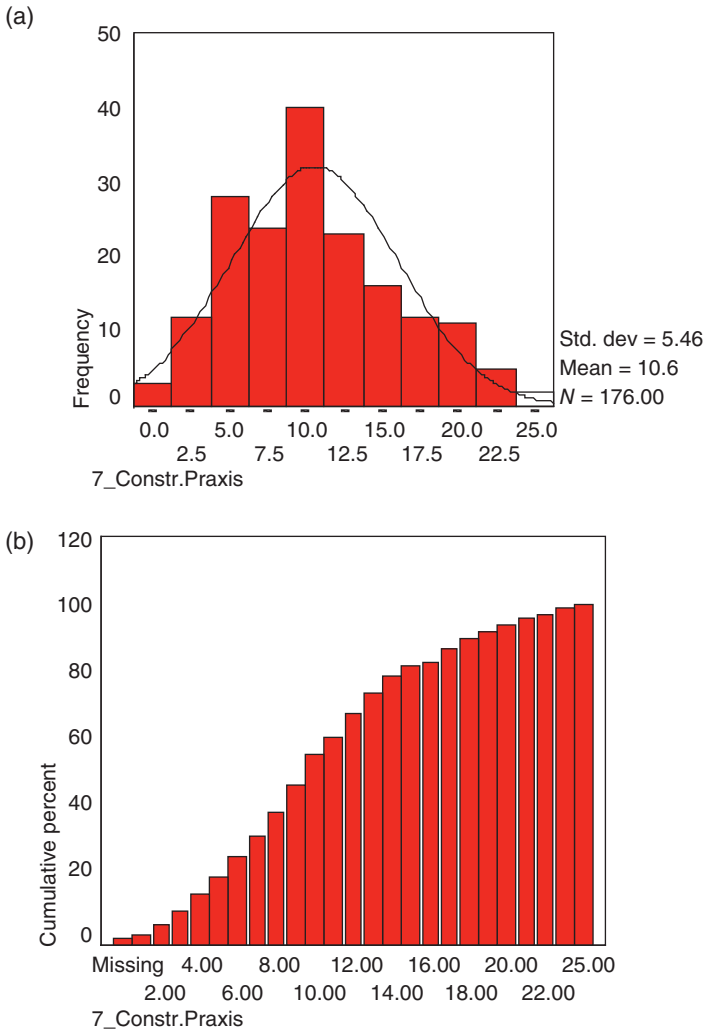


Figure 2.2 Empirical approximation to the "bell" normal curve and the ogive curve, illustrated with *C-Praxis* test data of AD patients from a study conducted by Dr. Erin Bigler in the Psychology Department at Brigham Young University. (a) Frequency/densities in intervals, bell curve. (b) Cumulatively added frequencies.

nonmathematician—as shown in the footnote.[2] However, the rationale of the raw score formula shown in footnote 2 can be converted into a simpler formula based upon Z-scores that approximates the results of the raw score formula and makes the rationale more clear.

$$p(z) = 0.4 \times 0.6^{(z^2)} \quad \text{(simplified approximation to the normal curve formula)}$$

This simple formula[3] may offer a sufficient approximation for many situations. To use the formula, we first transform the raw score values X into Z-score values to be used in the formula's

2 $P(X) = \left(1/\sqrt{2\pi\sigma^2}\right)e^{-(x-\mu)^2/2\sigma^2}$

3 In an earlier decade this rounded-constant formula, with the use of a small calculator, would have reduced the need for normal curve tables, but now spreadsheet commands are as available to most scholars as are calculators. However, practice with using this simplified formula has pedagogical value in demonstrating the parts and the rationale of the formula, as shown in what follows.

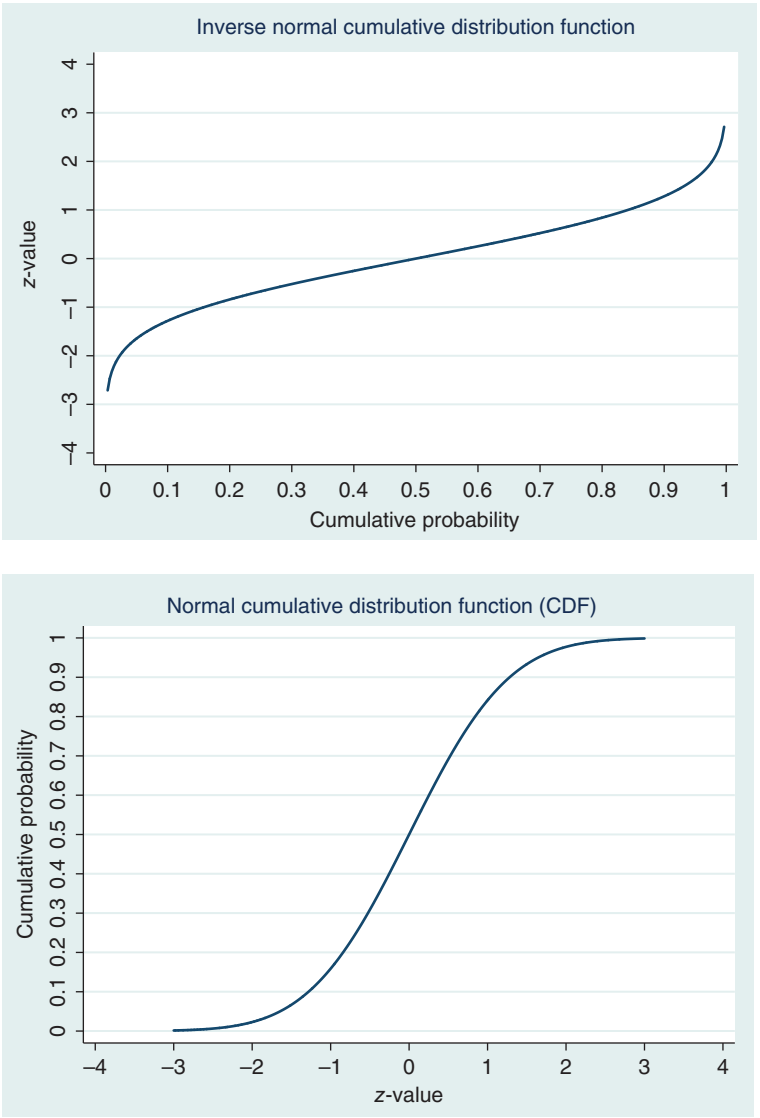


Figure 2.3 Inverted normal ogive curve above, compared to normal cumulative distribution function (CDF) ogive curve below.

exponent. The approximate formula is quite robust, even if the two numerical values in the formula (0.4 and 0.6) are rounded to one digit for simplicity and memorability, they will still create a normal curve comparable to the one created by the full formula of footnote 2 and embodied in Figure 2.4. For the example given in Subsection 2.1.2.3 of the 80-year-old woman, whose score of $X = 9$ on the *C-Praxis* test was one SD unit below its mean, $Z_x = -1.0$, we use the simplified approximation equation to find a probability density value, $p(Z)$, of approximately 0.240.

$$p(Z) = .4 \times .6^{(z^2)} = .4 \times .6^{(-1)^2} = .4 \times .6^1 = .240.$$

Similarly, for the 50-year-old man in that same section whose score of $X = 12$ on the *V-Fluency* test was an SD unit of 1.2 below its mean, $Z_y = -1.2$, we find a probability density value, *p(Z)*, of approximately .192.

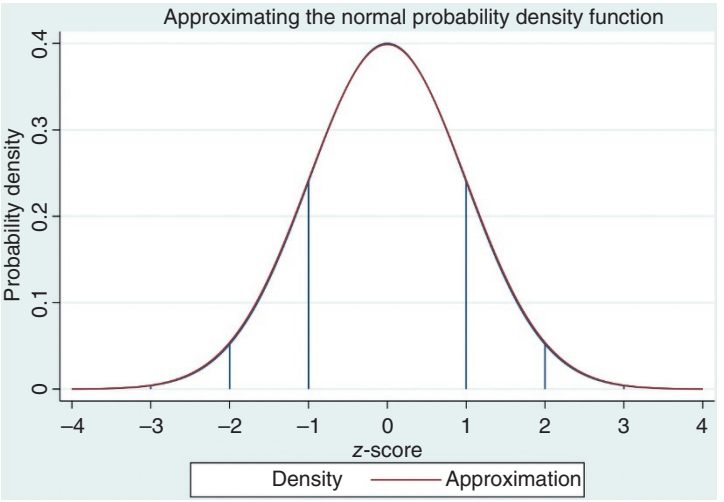


Figure 2.4 The theoretical law of the normal distribution, uses the formula in footnote 2, to calculate *probability densities* as a function of the number of SDs above or below the mean. This is referred to as a *standard normal curve* because it constructs a normal curve across the range of the *Z*-score values, that is, across the standardized values.

$$p(Z) = 0.4 \times 0.6^{(-1.2)^2} = 0.4 \times 0.6^{1.44} = 0.4 \times 0.4792 = 0.1917 \cong 0.192$$

The tabular formulas and values at the top of Figure 2.4 verify that the $p(Z)$ value corresponding to a *Z*-score of $Z = 1$ (or of $Z = -1$ for that matter) does indeed correspond to a *probability density* value of approximately 0.24. Also, looking at the line drawing of the bell curve we see that the vertical riser for a *Z*-score of −1.0 intersects the curve at that height, as shown on the probability density axis to the left. We also see that the vertical riser for a *Z*-score of 2.0 (either +2.0 or −2.0) intersects the curve at a height of about 0.05. There is no vertical riser for a *Z*-score at −1.2, but following the *Z*-score values on the abscissa of that figure and the probability densities along the ordinate, it can be observed that if a riser were placed at −1.2, it would intersect the curve a small distance below the probability density marker for 0.2, which corresponds to the obtained $p(Z)$ value of 0.192.

Standardized normal probability density				
Exact		**Approximate**		
$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$		$p(z) = 0.4 \times 0.6^{(z^2)}$		
Z	0	+1	+2	+3
$p(z)$	0.40	0.24	0.05	0.004

Probability density values with five decimal place accuracy (or greater) can be obtained using a spreadsheet. In Excel, for example, entering the command "**=norm.s.dist(–1.0, FALSE)**" returns

a value of 0.24197, which verifies reasonably well our obtained value of 0.24 from the approximate formula given at the top of Figure 2.4. Using a similar command with $Z = -1.2$, "**=norm.s.dist (–1.2, FALSE)**" returns a value of 0.19419. In other words, by changing the second argument of FALSE to TRUE in this command, Excel reports the area of the normal curve to the left of the *Z*-score rather than the probability density for that *Z*-score value. As another example, to calculate the area of the normal curve for *Z*-scores of less than +2.0, the command is "**=norm.s.dist (2, TRUE)**", which returns the value of 0.97725. Subtracting this from 1.00000, the total area under the normal curve, gives 0.02275. In other words, only about 2.3% of the area of the normal curve is above a positive *Z*-score value of 2, or below a negative *Z*-score value of −2.

This illustrates the importance and centrality of the SD in estimating and graphing statistical information. If the normal ***N****-law* holds for the data, that is, if the data distribution is close to being Gaussian, then the *SD contains all the information* necessary to derive the probability densities of point locations and also the percents (or proportions) of normal curve areas for desired *Z*-score intervals. For standardized data, that is *Z*-score data, the *Z*-value is the *only* parameter needed for the normal PDF $p(z)$.

2.1.4.1 The Rationale Underlying the Simple Approximate Probability Distribution Equation

Breaking down the standardized version of the probability density formula given at the top of Figure 2.4 makes the rationale of it and also the rationale of the more complex unstandardized formula more understandable. There are only three parameters from which probability density is calculated: first is the *Z*-score; the second is the constant pi, $\pi = 3.1416$; and the third is the constant $e = 2.71828$, the base of Napierian logarithms.

The first component in the simplifying formula for approximating the probability distribution function is based upon the constant pi (π). It is the inverse of the square root of pi multiplied by two: $\frac{1}{\sqrt{2\pi}} = \frac{1}{\sqrt{2(3.1416)}} = \frac{1}{\sqrt{2(3.1416)}} = \frac{1}{\sqrt{6.2832}} = \frac{1}{2.5066} = 0.3989$. This result is rounded to 0.4 for use in the approximation formula. The second component is *e,* the base of Napierian logarithms taken to the power of the negative of one-half: $e^{-\frac{1}{2}} = \frac{1}{\sqrt{2.71828}} = \frac{1}{1.6487} = 0.6065$. This is simplified for use in the formula by rounding to 0.6. The third component is the square of the *Z*-score value for which we are seeking the probability density, $p(Z)$. It becomes the exponent for the second component, which is then premultiplied by the first component to produce the approximation for the probability density for the *Z*-score value: $p(Z) = 0.4 \times 0.6^{1.44} = 0.4 \times 0.479 = 0.192$.

2.1.4.2 The Need for a Cumulative Density Function

Be careful not to misinterpret the probability density $p(X)$ to be the actual probability p for a score in the *X* distribution. Point values of continuous variables do not have probabilities. They have probability densities. Numerical intervals for these continuous values, however, do have probabilities. The curve of the ***N***-law only holds for continuous variables like those we are using here. This is an exception to our usual practice of using discrete integer data for numerical demonstrations. In the continuous case, the probabilities are calculated for a particular *interval of locations*. For that reason, we need the *cumulative density* function—in the empirical form of the step-like approximation to the normal ogive. For example, the "(b) cumulatively added frequencies" on the bottom graph of

Figure 2.2 arose by consecutive addition of frequencies under the "bell curve" in the individual locations with the summation going from left to right. Analogously, for truly continuous data, the cumulative density function given in footnote 2 is accomplished by consecutive *integration*, using integral calculus, of frequency densities.

2.1.5 (§18) The Ogive Curve or Normal Curve Cumulative Distribution Function *N(Z)*

The ideally smooth—continuous *normal ogive curve*—as quite closely approximated in the second graph of Figure 2.2 in Section §17—is then called the *normal curve cumulative distribution N*(*Z*). Unlike the stepwise adding of discrete interval frequencies, the continuous ogive arises by integrating the density function $p(z)$. The integral expresses the probability $P(.)$ of finding a certain unknown value Z in the infinite interval from minus infinity to another given value of Z, or roughly, from the approximate minimum around -3 up to the given Z. For instance, we already know that $N(Z = 0)$ has a 50% probability, $P = 0.50$. It is the probability that some yet unknown Z-score of a certain AD patient will be found somewhere between minus infinity and the mean, practically between the approximate minimum $-3z$ and $Z = 0$. Of course, this is because the normal distribution is symmetric, and the mean is identical to the median, becoming the 50th percentile. Thus, for the probability of a score being below the mean, the probability of $Z \leq 0$, is $P(Z \leq 0) = N(Z = 0) = 0.50$. Similarly, we can find from the normal ogive that $P(Z \leq 1) = N(Z = 1) = 0.84$. However, the probability of the occurrence of a score in a finite interval, say between $Z_1 = 0$ and $Z_2 = 1$, is then defined as the difference between the two function levels:

$$P(Z \text{ is between } Z_1 \text{ and } Z_2) = N(Z = 1) - N(Z = 0) = 0.84 - 0.50 = 0.34.$$

It represents the probability that a score will fall into the interval between the mean ($Z = 0$) and 1 SD above the mean. This is expressed as the area under the bell curve between the two vertical risers in Figure 2.5 and/or as the difference between two values N_2 and N_1 in Figure 2.6.

The "rule of ±3 SD" of Section §16 was derived in this way. Usually, and especially in smaller samples, we are satisfied with the use of only ±2 SD and the corresponding probability of 0.95 for the estimation of the minimum, maximum, and range of the given sample. Besides the several indicated values, there are also a couple of further values of the probability integral $N(z)$ to be kept in mind by a researcher and clinician—as will be shown. But what about the other values? They can be found in statistical tables, but a more practical way to find these values would be to use a simplified equation similar to the *probability density approximation equation*.

While the computation we recommended for the density function $p(Z)$ was not exact due to the rounding errors involved, we now suggest approximating the integral[4] $N(Z)$ by simulating it in terms of a different function. Its values agree with the true $N(Z)$ again up to two decimals. Such an approximation or rough empirical "almost-equivalency" is not rare in science. For instance, compare the description of a learning curve in terms of an exponential as well as a parabolic function. It is the congruency of a function with another theoretical network of interrelated laws that suggests its preference for its competitors, which can yield roughly similar values derived from a different mathematical background. In the case we suggested here, however, the curve simulating the normal ogive has its origin in the psychometric theory of Item Response Models. Around 1960, it was suggested by Allan Birnbaum, and recommended by Lord and Novick (1968), to

4 $N(z) = \int_{-\infty}^{z} p(z)dz$, cf. Figures 2.4 and 2.5.

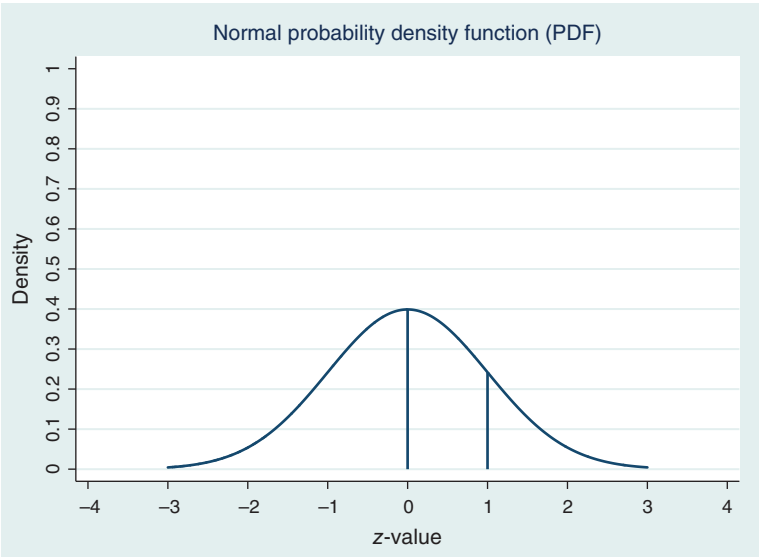


Figure 2.5 Probability (*P*) of *Z*-scores between 0 and 1 shown as the area under the normal curve between the two vertical risers. Found by integrating the probability densities between these two scores. Also shown as a normal ogive in Figure 2.6 for comparison.

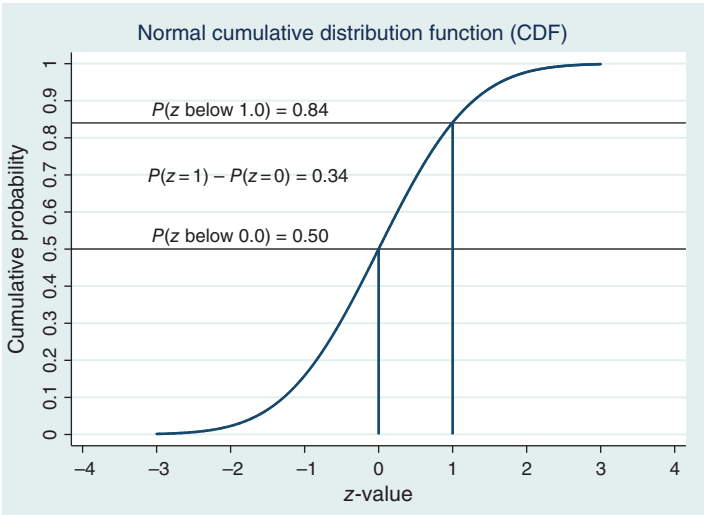


Figure 2.6 Normal ogive curve, the cumulative normal distribution function found by integrating the probability densities from left to right under the bell curve of Figure 2.5.

roughly approximate the normal ogive with sufficient practical usefulness by a function generally known as the *logistic*, and referred to by Birnbaum as the "logistic cumulative distribution function":[5]

5 Equation 17.2.1 in chapter 17 of part 5 (contributed by Allan Birnbaum) of the Lord, Novick & Birnbaum (1968) book.

$$N(z) \approx \frac{1}{1+e^{-1.7z}} \cong \frac{1}{1+2.72^{-1.7z}}.$$

The value $N(Z=0)=0.50$ is generated exactly: $1/(1+0.272^0)=1/(1+1)=0.50$. Or repeating the aforementioned question "what proportion of people in a normally distributed population is below $Z=1$," we may use a calculator with an *e*-key to easily[6] find the value $N(Z=1)=0.84$ to be approximated by 0.845 (while the still more precise value in statistical tables is 0.841), etc. Most of the other values if rounded to two decimals would match the exact values in tables quite well—i.e. if multiplied by 100 to express them in percents, they mostly agree up to 1%.

2.1.6 (§19) Percentile Position of a Score and the *N*-Law of Distribution

In our former illustration of "how to compare the incomparable," we saw that by shifting the origins of scales to their means and taking the SDs as new units of measurement we can compare two quite distinct values according to their relative position with respect to their means and with respect to their distance from the means. This was to compare them relatively with respect to the variability of the samples they were coming from.

A similar idea would be to compare the two subjects' scores—relatively again—with respect to the proportion of people who are better and/or worse in their samples or in the corresponding categories by age, sex, education, etc. For example, if a 50-year-old man's performance in the *C-Praxis* test is better than 90% of the other men of his age category, it means that he is relatively better than an 80-year-old woman who, on a quite different test of *V-Fluency*, left only 70% of her category mates behind.

However, if the score values follow the ***N***-law of distribution then the percentage under a given score is completely established by its corresponding number of SDs below or above the mean. It is completely specified by its standardized *Z*-score points, just through the function $N(Z)$. For example, we need not order several hundreds of data points by their magnitude into a list and count how many people are below a given score *X*. Having known that the score *X* has its standardized value of $Z=1.0$, we can use the simple *logistic* equation given in Section §18 to find out that $N(z=1)=0.84$, that is, 84% of the people from that category have a lower score. Moreover, we say that the standardized *Z*-score of 1.0 is the 84th *percentile*. Of course, the value is accurate up to the degree to which the distribution of the given data approximates the theoretical ***N***-law.

Using the formula above for computing $N(Z)$, we can establish other percentiles. For example, we can verify that the standardized score $z \cong 2$ is at the 97.5th percentile, etc. But some other useful values that are often used in practice are usually related *not* to the original value of $N(Z)$ but rather to their differences corresponding to certain *Z*-values, i.e. to SDs around the mean. Typically, you should remember that for certain symmetric intervals such as that between $z=-1$ and $z=+1$, there is a proportion approximately approaching the probability $P=N(+1)-N(-1)=0.68$.

Expressed in percents, other common ones are:

Between the *Z*-values or SDs about the mean				Approximate proportion
−1	to	+1	...	68%
−2	to	+2	...	95%
−3	to	+3	...	>99%

6 As $Z=1$ we seek $1/(1+e^{-1.7\times1.0})=1/(1+e^{-1.7})$. A calculator gives us $e^{-1.7}\cong0.183$, then $1/1.183\cong0.845$.

Since the normal distribution is symmetric, we can reconstruct the formerly known values of $N(Z)$ from the following ones that are recommended for remembering:

$N(z = 1) = 0.84$ ... divide 0.68 by two due to symmetry and add 0.50 for below the mean,
$N(z = 2) = 0.975$ ... divide 0.95 by two due to symmetry and add 0.50 for below the mean

For example, the answer to the following question can be used as a model for similar problems:

Q: What is the proportion of subjects between +1 and −2 SD, i.e. between the Z-scores −2 and +1?

A: Since between the mean and –2, there is $0.95/2 = 0.475$, and between the mean and +1, there is $0.68/2 = 0.34$, altogether it is $0.475 + 0.34 = 0.815$, or 81.5%.

And so on, using simple calculations of $N(Z)$ for Z-scores outside the −3 to +3 range.

2.1.7 (§20) Are *Z*-Scores and Percentiles Fair?

To initiate critical thinking, suppose we give the following answer to the question that is the title of this section:

> No. Although Z-scores are quite good for cross-comparisons between "incomparable" scales, the Z-scores and the associated percentiles are not entirely "fair" for intraindividual as well as interindividual comparisons along the same scale.

Moreover, a little strangely, the reason for the "unfairness" is partly embedded in the N-law of "normality" itself. In part, it also depends on our intuitive sense of fairness of assessment, such as in the case of subjective feelings about the fairness of school grades.

Notice that Z-scores and percentiles are mutually related, and the relationship is *nonlinear*. This is clear, of course, because the normal distribution function $N(Z)$, the ogive, is not a straight line. On the other hand, we have described the standardization of the original *raw score X* into a *Z-score* as translating its scale origin to the mean, and substituting the original scale unit by the length of the SD—this is known as a case of a *linear transformation*. Let's ask about the practical consequences and possible interpretations of those two transformations. Let's also ask: are they "fair" to all individuals or should they be modified to make them more intuitively appropriate?

First, let's have a look at some practical consequences of the fact that the Z-score is a linear transformation of the original raw scores, say X or Y, from the point of view of the intraindividual *change of performance* on a psychological test. Let's take our former *V-Fluency* test with its mean and SD as known from the reference normative study by Welsh et al. (1994): $\overline{Y} = 18$[words], $s_y = 5$ [words], and the standardization formula to transform the raw Y-score into the Z-score $Z = (Y - \overline{Y})/s_y = (Y - 18)/5$. If an average individual improved his/her performance from the mean score 18 to the improved value 23, i.e. producing five more words, he/she will be "rewarded" by the corresponding Z-score increment 1, since that subject has moved from $Z = 0$ to $Z = 1$. The same Z-score increment, i.e. +1 will be the "reward" for a further improvement of five words from 23 to 28, and the same for the improvement from 28 to 33. The last one expressed in Z-score form is from 2 to 3 having reached the approximate maximum of words an individual of that age category can produce in one minute.

But is that "fair"? Is an increment of performance that is close to the mean equally as "difficult" as the same improvement that is close to the maximum or "ceiling?" Still more illustrative is the improvement of performance in the high jump from 100 to 105 cm compared to an improvement from 200 to 205 cm. Both of these improvements in athletic performance will be "rewarded" by the

same increment after their transformation into *Z*-scores. And this kind of "injustice" also holds for interindividual comparisons of the two individuals' increments in the frame of the same sample or population.

Intuitively, we feel that improvements close to the maximum should be "rewarded" more than the same ones close to the average. And vice versa: the decrements close to the minimum should be "punished" by larger decrements on a point scale in a "fair" assessment. Therefore, we feel that the points of the new and "fair" scale should not be a linear transformation of the raw score, or consequently, of the *Z*-scores either.

As we know the percentiles are produced by a nonlinear transformation through the normal ogive function $N(Z)$, let's see if they are "better" from that point of view:

Changing the *Z*-score from:	Results in a shift from the ...	To the ... percentile	Percentile change of
0–1	50th	84th	34.0
1–2	84th	97.5th	13.6
2–3	97.5th	99.9th	2.1

Thus, the transformation into a percentile scale is not better, it is much more "unfair." Actually, its curvilinearity works in an opposite direction than we would like. This can be clearly seen in Figure 2.3. The middle part of the ogive is almost a straight line but to correct the "unfairness" of that linearity, we would need to have its ends bent in the opposite direction. To accelerate the "rewards" and/or "punishments" for extraordinarily good and/or bad performances, we would need a curve in the general form of an anti-ogive.

An appropriate and "fair" way to better reward the excellent raw scores or *Z*-scores can be based on different *theoretical* or *practical* considerations that would suggest which specific nonlinear functions to use. Certain approaches can be found in practice that are intentionally or unintentionally based on the logarithmic Stimulus–Reaction (*S*–*R*) law or Weber-Fechner law well-known from the early beginnings of psychophysics:

> On lower levels, the increments of reactions *R* produced by increments of stimuli *S* grow linearly but for more and more intensive stimuli *S*, the *R*s seem to approach a ceiling and even much larger *S*-increments produce very small or no *R*-increments.

But, very often, there are no theoretical reasons for establishing a *nonlinear transformation function* as in the case of the logarithmic relation between *S* and *R*. Then, why not use the ***N***-law again, but this time in an opposite or inverse way. The original nonlinearity of the ***N***-ogive is suppressing the differences of transformed scores in a certain "unfair" way, even more "unfair" than the linear *Z*-scores.

2.2 Combining and Using Standard Scores

Section §21 shows how *Z*-scores can be transformed to alternate scales with a variety of different means and SDs. Such alternatives are sometimes preferred in order to have scores with positive integer values for clinical or educational uses. In Section §22, we consider cases where the distributions of variables depart from normality or distributional symmetry through traditional skewness or kurtosis, the influence of floor or ceiling effects, or more extreme J-shaped or L-shaped distributions. Some variables naturally have nonnormal distributions, and non-normal distributions can

often be expected when extreme groups or unusual subpopulations are selected for analysis. Section §23 addresses the somewhat controversial idea of creating a composite score by taking the sum or mean of the standardized Z-scores for multiple variables. Such an "overall" measure implicitly weights each of the components equally, without addressing their empirical intercorrelations or theoretical interconnections. Section §24 discusses various ways of weighting subtests and strategies of compensation.

2.2.1 (§21) More Convenient Scores Derived from *Z*: The Deviation *IQ*, *T*-Score, and Stanine

There are several ways of scoring, which represent simple transformations of the original Z-standardization. One motivation for preferring them to the original Z-score scale is for clarity, another is convenience. Maybe for a vocational adviser, it is not too helpful to explain to a client that his/her diagnostic result is 2.83, which means he/she is exhibiting a very good performance in an admission test and/or that the score −1.09 is rather not too good.

Scale	Mean	Standard deviation	Expected range
T-scores	50	10	20–80
Deviation *IQ* scores	100	15	55–145
Standard nine (stanine) scores	5	2	1–9
Standard 10 (sten) scores or deciles	5.5	2	1–10
SAT/GRE scores	500	100	200–800
ACT scores	18	6	0–36
Normal curve equivalent (NCE) scores	50	21.063	1–99

"Inconvenient" distributions such as these can be easily transformed by shifting the scale so that the mean, the $Z = 0$ point, is given a more favorable value, say 100. And the SD, the $Z = 1$ scale unit, is assigned a new number such as 15. Then a subject with a score located at 1 SD below the mean, i.e. with $Z = -1.0$, will now be assigned 85 points. This is how the Wechsler Intelligence Scale results are scored. This common way of expressing an Intelligence Quotient in contemporary times is called the *Deviation IQ*, and it is computed from Z-scores by the linear transformation indicated earlier, *Deviation IQ* $= 100 + 15Z$, whose concrete value shows the individual's deviation from the average "norm" of 100 points. Apparently, the shift made by adding 100 was chosen to follow the historical age-based IQ: *Mental Age* divided by *Chronological Age* multiplied by 100, which yields a score of 100 if the Mental Age corresponds to the actual age of the subject. The number 15 was established rather deliberately to cover the usual extreme values of the age-based IQ in the past. Thus, a usual top level of the Deviation *IQ*, the $+3Z$, now becomes $100 + (3 \times 15) = 145$, which seems a reasonable approximate maximum to be exceeded only in the case of really exceptional people. Also, of course, the scores are rounded to the nearest integer to avoid inconvenient decimal digits.

Another Z-score transformation, which is used very often, is $T = 50 + 10z$. The multiplication by 10 and subsequent rounding is to handle the inconvenient decimal values again. Shifting the mean to 50 avoids negative values, and the SD is assigned to be 10 points. Only in rare cases do the T-points fall outside the range from the minimum 20 and the maximum 80 due to possible non-normality of the distribution. This T-scale scoring is traditionally used by the well-known *MMPI* personality test, for example.

Although the *Deviation IQ* and *T* transformations use rounded values, their scales still have quite high resolution. Several rather rough scales still in clinical use profoundly reduce the variance in measurement values with the attendant reduction in "discriminability" and validity as mentioned earlier. One of these ways of scoring is the *Stanine* scale, named after its STAndard NINE points covering the range according to *Stanine points* $=5+2z$. This means that the classical SD standard *Z*-scores are transformed and rounded into a 9-category scale with a mean of 5. It also clips the tails by assigning observations above 2 SD to the maximum 9 points and under −2 SD to the minimum 1 point.

2.2.2 (§22) What If the Scores Do Not Follow the *N*-Law? A Skewed Bell Curve

The *T*-score scale is very often presented to the public as a scale from a minimum level of 0 to its maximum at 100, but with an observation that those extreme points are not usually reached. As we know, due to the ***N***-law and its rule of 6 SD, the range of the *T*-scale should be expected from $-3z$ to $+3z$, which yields a range from 20 to 80 for the *T*-score, according to the formula. Then, why the presentation of the 0 to 100 range? Is that really for the layman's convenience? Similarly, the *Stanine* scale treats the possible extreme raw scores by neglecting distinctions among those that exceed the ±2 SD interval. Again, is there not another profound reason for doing this?

Although some people believe that the ***N***-law is universal in all of nature, this is not the case. At least some psychological tests, for instance, the tasks that depend on persistence, are among those that exhibit a "distorted bell" curve—the *skewed* or *asymmetric* one. It covers general endurance activities, too. This can be observed quite easily just by watching marathon runners as they are caught on camera from a helicopter in an online TV transmission during a competition. At the very moment the winner is known, the main body of the runners and their "density" is "distributed" along the road with a somewhat bell-shaped curve except for the long and sparse "tail" of the curve consisting of the last runners. This right-skewed distribution for the Boston Marathon curve is shown superimposed upon a normal curve in Figure 2.7 for contrast.

A similar skewed distribution from a large database gathered from public websites from 1970 to 2013 is shown in Figure 2.8. This large database was assembled from many publically available

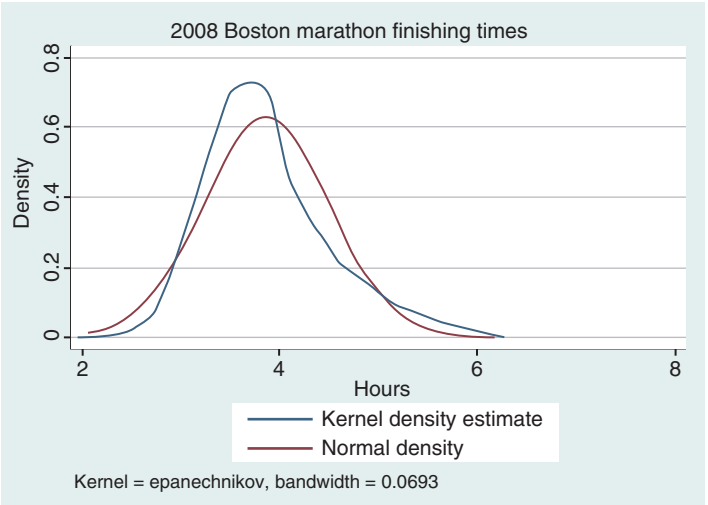
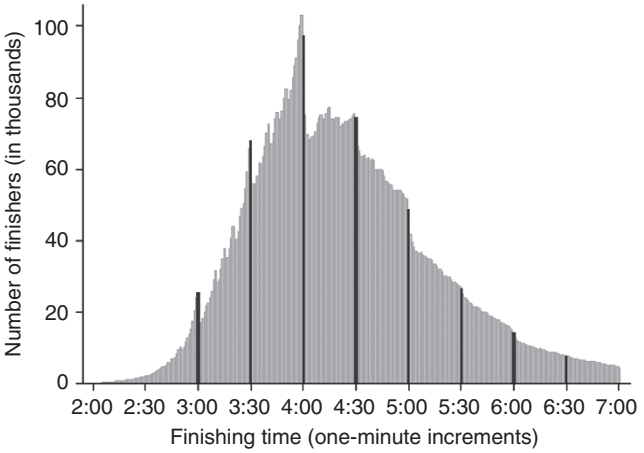


Figure 2.7 Asymmetric frequency curve of finishing times for runners in a marathon.



Note. The dark bars highlight the density in the 1-minute bin just before each 30-minute threshold.

Figure 2.8 Demonstration of the reference-dependent preference effect on finishing times cumulated over n = 9,789,093 marathon runners. *Source:* Allen et al., 2017 / INFORMS.

datasets of marathon finishing times, gathered from 1970 to 2013. Allen, Dechow, Pope, & Wu (2017) assembled this database for the purpose of testing the hypothesis, based upon prospect theory,[7] that "reference-dependent preferences" affect marathon finishing times. This particular bar graph in Figure 2.8 is quite different from the smoothed line graph in Figure 2.7. The most obvious difference is that Figure 2.8 is a bar chart with 300 gray bars (one for each of the 300 minutes in the five hours) rather than a smoothed line chart. It is also punctuated with black reference point lines marking off every 30 bars, that is, every half hour. The authors hypothesized discontinuities in the curve at the reference points, particularly for those at three, four, five, or six-hour points, with more finishers than usual finishing before the major reference point and fewer after. This is seen somewhat at the three-hour mark, but especially at the four-hour mark. It is interesting to note that the mode of the distribution is observed to be the grey bar that extends from 3 : 59 to 4 : 00, that is, the last minute before the four-hour reference point.

One of the lessons from Figure 2.8 is that while psychometricians may perceive irregularities in data distribution shape as an irritation and a problem to be solved, they may present more possibilities than problems. They may, in fact, present opportunities for sociological and psychological "natural experiments" of the kind reported in the Allen et al. study. They could become fertile occasions for creating potentially high-impact studies with datasets from various public websites such as NHANES.[8]

7 The foremost account of reference-dependent preferences is prospect theory (Tversky & Kahneman, 1974, 1991; Kahneman & Tversky, 1979). According to prospect theory, a reference point determines how an outcome is perceived, and the reference point in this study is "finishing time as a round number." The specific hypothesis is that perhaps runners unconsciously adjust and manage effort in accord with whole number hours. That is, the outcome of one's performance in a marathon is evaluated more favorably if the finishing time is 3 : 59, that is three hours and 59 minutes, rather than 4 : 01.

8 NHANES is the acronym for *National Health and Nutrition Examination Survey*. It is affiliated with the *National Center for Health Statistics*, a division of *Centers for Disease Conrol and Prefention* (CDC). https://www.cdc.gov/nchs/nhanes/index.htm.

2.2.2.1 Floor and Ceiling Effects: J-Shaped and L-Shaped Distributions

We return now to the problem of how to manage irregular data, that is, "scores that do not follow the *N*-Law." When the asymmetry is negligible, the ordinary *Z*-scores, *T*-scale scores, and even *Stanine* scores may still be used, but with a certain caution. The 0–100 range of *T*-scale scores provides a certain reserve or buffer zone so that the minimum and maximum of the original raw scores would not exceed the range of *T*. In the marathon run example, that could happen only if the right tail of the curve were so long that the maximum time score was over 5 SD above the mean which would be very extreme asymmetry. Of course, in similar extreme cases of departures from the normal distribution, the mean and SD can become quite inappropriate statistical characteristics.

Besides the performances that depend on persistence or endurance, there are many other types of tasks whose fulfillment is limited in a certain sense by a "ceiling" or "floor" or both. For instance, one of the tests for AD patients is the Trail Making Test. The patient is asked to draw a "trail" (line), which connects scattered points to create a series in an appropriate order. The test score is timed in seconds, and it expresses the patient's speed, but another by-product score is the number of mistakes. If the sample of subjects performs very well—or if the test is rather easy for them—the main body of the subjects, i.e. the mode Mo( ) of the time score, will be very close to the left "ceiling" of almost zero time as in the hypothetical curve in Figure 2.9. Actually, there are two "ceilings" here. One is the theoretically shortest time of best human performance in the test. The second one is the physically possible shortest time close to absolute zero time. An extreme asymmetry may be caused by the fact that only one half of the "bell curve" is found in the sample, either the J-shaped (left) tail or the L-shaped (right) tail.

There are several types of distribution functions that are non-normal due to the substantive processes producing the stochastic phenomena. These could be such things as physical, mechanical, or socioeconomic processes. Thus, we can find some special distributions for very rare events like the Weibull or Poisson distribution. The latter describes the probability distribution of a dichotomous variable (with outcomes 0; 1) when one of its outcomes is extremely rare in a very large sample. As the frequencies of the rare event increase, the Poisson distribution converges to the binomial

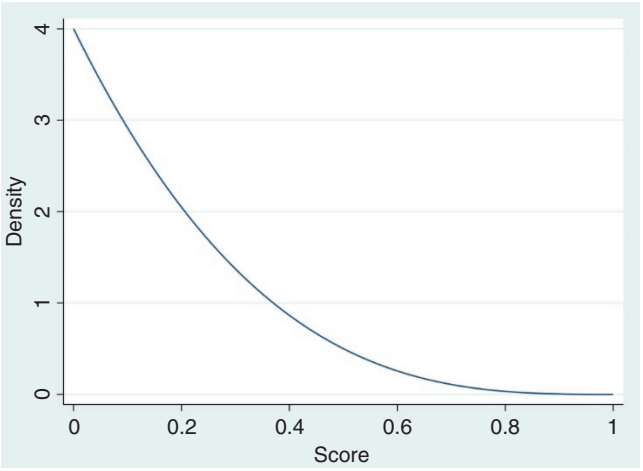


Figure 2.9 An illustration of extreme asymmetry: the *L*-curve distribution.

(approximately normal) distribution. However applicable this distribution may seem, it is not typical of dichotomous items. Items with extremely small p, and/or q, have very small variance, hence a low "discriminability." Therefore, their validity for predicting any criterion will be low, rendering the items nearly useless.

In behavioral science, most non-normal distributions can be assumed to depart from the normal due to some external influence, for instance, due to the ceiling/floor effects mentioned earlier. One of them may be due to the manmade *construction* of the diagnostic procedure. In the case of performance tests, the test task can be too difficult/easy, meaning too close to the limiting value of a ceiling/floor. If, for example, a Trail Making Test item had been particularly difficult, thus creating a longer trail to be made by a tested individual, then the average performance time would also have been longer, that is, located farther from zero on the scale. Hence, the left-side squeezing effect of the physically impossible performance better than 0 [seconds] might become negligible. In such a case, it is better to modify the construction of the diagnostic method to tailor it for the desired practical purpose and a given population.

2.2.2.2 Variables with Inherently Non-normal Distributions

Another type of non-normality can be due to the *natural* character of the variable measured. For example, the body weight in normal populations is known to be a little asymmetric. Sometimes it can be transformable into an approximately normal distribution by taking logarithms of the measured values, which squeezes the scale x. With unequal scale intervals of $\log x$, the curve becomes a symmetric "bell," again. It may suggest that perhaps the genuine "naturally normal" units of measurement should be logarithms of the units used, not the original units themselves. Such a distribution is known as *lognormal*. Some people seem to believe that with appropriate scaling all variables in the universe have to have a normal distribution. If they were right, we would hardly find nonlinear relationships in nature.

In usual behavioral research, there might be some rather rare cases when one may take a certain specific form of transformation as given. For example, if the logarithmic transformations earlier were found generally useful then it would be meaningful to use them everywhere as a standard all the time. This is a type of monotone transformation, meaning that with growing x, its transformed values also grow. However, some strong advocates of the "normal distribution everywhere" recommend doing almost anything to make the distribution normal, including the use of nonmonotone transformations! Moreover, some suggest taking the observed frequencies on intervals of the measured variable, then finding their corresponding intervals on the standardized Z-scale in the table of normal distribution probabilities, and then forcing these normal intervals into the original scale units. This method used for an ordinal scale seemingly makes a forced "rectification of ordinal data into an interval scale." Say, some grades on a Likert scale like 1; 2; 3; 4; 5 can be purposely compressed, some can be extended, and finally, the resulting scale would have a normal distribution, but now with scale values transformed as e.g. 1; 2.1; 3.8; 4.2; 6.4, and so on. In this way, it is possible to deliberately force any kind of distribution into a normal one, for example, even bimodal (with two "humps") distributions, which arose from an inappropriate heterogeneous mixing of two populations.

However, there is a big problem with this way of trying to make complicated things simple. First, in general, we may ask whether it is really *normal* to assume that there is only one type of distribution function, namely the "normal," strictly everywhere in the whole of nature and the cosmos. Second, the assumption of "normality everywhere" produces a consequence of "linearity everywhere" (with some special exemptions). Such an assumption—that (almost)

anytime you find two variables and put them together, they should be related just by a linear regression—is certainly untrue. In that case, all nonlinear functions of mathematics would be useless for any practical application.

2.2.2.3 Population Selection and Extreme Groups

Selecting a special subpopulation can be another source of asymmetry in probability distributions. The selection can be due to a specific interest in a given research study or because of *self-selection* by the research subjects themselves. This can be seen, for instance, when unqualified candidates or others, apply for admission to certain institutions, organizations, etc., especially if it is automatically assumed that only top-level people are eligible. Another example can be research focused on a subpopulation created by a certain *cut-off score*, i.e. just on the people contained in the left and/or right *tail of the distribution*. For instance, the study may be focused only on especially gifted children, or the aim of the research may be to compare the psychological characteristics of extremely obese people with those in the opposite tail of the normal distribution. Figure 2.9 in Subsection 2.2.2.1 illustrates how the distribution of the people above a cut-off score represents such a tail. In such and similar cases, it would be contradictory to use statistical methods that are based on the assumption of normality. Then, typically, such procedures for computing intervals of estimation, tests of hypotheses, etc., are not usable. In such situations, it is necessary to use distribution-free methods, nonparametric methods.

A specific kind of non-normality arises when two or more variables are joined by a computational formula to yield a new variable, a kind of joint "index" or so. The problem consists in the fact that if the two (or more) original variables were distributed normally, then it holds that *a nonlinear function of normal variables has a non-normal distribution*. Linear functions of a variable like taking the sum, mean, or multiplication by a constant will leave the distribution of the new variable normal. However, a transformation like second or higher powers, or the square root, will produce a variable that will not be normal—simply by definition. Some additional examples of how to create a non-normal variable from the normal ones:

> *Ratios of two* normally distributed variables are usually used as favorite indices in medicine, namely if any variable is "objectivized" or "relativized" by expressing its value with respect to "1[kg] of body weight" or so. It follows that any such ratio does not have a normal distribution. For instance, the "relativized" maximal volume of consumption of oxygen, used as a physiological measure of endurance: VO_{2MAX} [ml]/Weight [kg]. Another example is the familiar Body Mass Index (*BMI*), which is computed as Weight [kg], (x) divided by squared Height [m], (y), i.e. $BMI = x/y^2$, where the distribution of each separate variable is generally supposed to be normal or close to normality.

Unfortunately, the problem is that such intentionally constructed variables with the explicitly built-in non-normality created by the researcher are then very often treated by methods that are explicitly based on the assumption of normal distributions.

2.2.3 (§23) Composite *Z*-Scores of Different Tests: Adding Apples and Oranges?

Similar to "comparing the incomparable," there is a way to sum up the different psychological test scores whose arithmetic addition otherwise would be meaningless. Once the raw scores have been transformed into *Z*-scores, *T*-scores, etc., they are "unit-less"—and they are "allowed" to be summed up together. For instance, the CERAD battery consists of tests that have test scores

measured in different units—seconds, number of completed pictures, number of mistakes, etc. If we would like to express the overall success of a certain subject on the whole battery, we need to transform the subject's scores from each of the tests into *Z*-scores, and then compute an overall *composite score* Z_c as a sum—or as an average—of n subtest scores for that individual:

$$Z_c = (Z_1 + Z_2 + Z_3 + ... + Z_n) = \frac{1}{n}\sum_{j=1}^{n} Z_j, \quad \text{or more simply,} \quad Z_c = \frac{1}{n}\sum Z_j\,.$$

One of the practical aims of the CERAD battery is to relate its scores to the general diagnosis of Alzheimer's Disease. Of course, using the entire battery offers more information than taking individual subtests separately one at a time. To evaluate the overall diagnostic quality of the battery, we might think about comparing or relating it to another usual diagnostic method. Such a method, for instance, is the MMSE, the Mini-Mental State Examination, a standard clinical method. As we will see later, a structural analysis of the CERAD battery shows that its subtests cover three different aspects of the "overall mental state." In short, these are "executive," "verbal," and "memory" aspects or kinds of *abilities*—so to speak. Therefore, to simplify the problem, we may wish to select just one subtest to represent each of the aspects, for example:

x_1...	C-Praxis test for "executive" aspect/ability,
x_2...	V-Fluency test for "verbal" aspect/ability, and
x_3...	L-Memory test for the "memory" aspect/ability.

In addition, we denote

y...	the MMSE result of an overall clinical examination.

The following data are rounded but real values of the means and SDs found in Utah patients (note that their mean level of performance in the tests is lower than in the Welsh et al. 1994 study):

Mean	SD	Units on the scale of scores	Scale's theoretical		Direction of the scale for a "good performance"
			Minimum	Maximum	
$\bar{x}_1 = 8$	$s_1 = 2$	No. of drawings copied	0	11	+
$\bar{x}_2 = 12$	$s_2 = 4$	No. of names of animals	0	∞	+
$\bar{x}_3 = 4$	$s_3 = 2$	No. of parts of a sentence	0	10	+
$\bar{y} = 20$	$s_y = 5$	MMSE points	0	30	+

+ *The higher the score, the better the performance.*

Those means and SDs can now be used to standardize the raw scores x_1, x_2, and x_3 of a certain AD patient by transforming his/her raw scores into *Z*-scores, *T*-points, Stanines, and so on. Let us use just the *Z*-scores. In a rather typical Utah male patient, we find his raw scores:

$$x_1 = 10, x_2 = 10, x_3 = 5, \text{which by our well} - \text{known transform } z_x = (x - \bar{x})/s_x \text{ become}$$

$z_1 = 1, z_2 = -0.5, z_3 = 0.5$, in *Z*-scores. If we would like to evaluate the relationship between the three subtests (x_1, x_2, x_3) on one side and the MMSE examination (y) on the other side, we can try to construct a kind of agglomerated score—*composite score*—of the former three. At first glance the summing up of their Z-scores seems to be an obviously simple way.

However, before summing up these Z-scores to get the patient's overall composite score, we have to pay attention to the separate test scales with respect to one another and to the real world. This includes such things as *meaningful scale direction*, and the correspondence of the separate test scales to the attribute being measured and its ideal concept of "best performance." In this example, each of the component scores is in the same positive direction in relation to that concept. So, the + and − signs attached to the computed Z-scores are correct and meaningful: "the higher the numerical score, the better the performance," e.g. here the *negative* $z_2 = -0.5$ really means a performance of half a standard deviation (SD) *below* the mean (the worst score of the three tests for that patient). However, for instance, imagine that if one of the subtests were measured as the time needed to complete a task, then we would have faced scores with an opposite scale direction: "the higher the time score, the worse the performance." Then the minus and plus signs of the Z-scores should have been inverted or reverse-scored before their addition. Therefore, before creating a composite score from several components or subtests, a careful "real-world" or common-sense evaluation of the meaningful directions of the separate scales is very important.

Now, using a formula for the average composite Z-score for $n = 3$ subtests, we get:

$$z_c = \frac{1}{n}\sum z = \frac{1}{3}(z_1 + z_2 + z_3)$$
$$= \frac{1}{3}[1 + (-0.5) + 0.5] = \frac{1}{3} \approx 0.33.$$

This means that the patient is slightly above average within his sample in the overall agglomerate evaluation of the three subtests.

2.2.4 (§24) Mutual Compensation, and Differential Importance of Sub-tests in the Composite

Among other problems of constructing an additive composite score, there are two main questionable properties: (i) The compensation of a low score on one subtest by a high score on another, and (ii) Summing the individual tests with implicitly equal weights of importance.

Compensation is formally due to the principle of summation and deals with all kinds of composite scores based on that principle. (There exist certain other ways to construct an overall result, but they are rather rarely used.) It may not be reasonable to consider that a complete absence of one component, such as an ability, could be fully compensated for by the others. However, the summation formula formally implies that. Considering not such an extreme case as the total absence of a component, note that the above composite score value, $z_c = 0.33$, may be a result of infinitely many combinations of its three component scores. While that patient was found above average in the executive test ($z_1 = 1$), and in the memory test ($z_3 = 0.5$), he was below average in the verbal test ($z_2 = -0.5$).

One simple change that does not influence the resulting composite $z_c = 0.33$ could be done through an interchange of signs + and − in the memory and verbal subtests. But are those two different sets of subtest scores with an identical composite score equivalent? Are the clinical conclusions about two such different patients the same? Another obvious example may be in diagnosing abilities for a job. The complete absence of politeness in a shop assistant cannot be fully compensated for by any other ability or skill he/she may possess, which otherwise are substantial co-agents for that kind of employment. For example, athletic performance in a certain discipline depends on a combination of several motor abilities such as endurance, coordination, and so on. Although these components of performance are combined in slightly different proportions in every individual, they cannot be combined without limits: In the long jump, a slower running phase cannot be compensated for by the explosive power in the moment of take-off. However, a

"compensatory model," such as that based on a simple sum of the components, cannot represent this fact if no additional restrictions are explicitly imposed.

In some cases, the problem can be partly overcome by using graphical *test profiles*. There are two main features of a profile of scores: (i) *Extreme limits* on some subtest scores, for instance, only a predetermined minimum score is allowed on a certain subtest, and a maximum on another, and (ii) *The overall form* of the profile, for example, scores in certain subtests should always be higher than in certain other subtests, showing the desired proportions of the composing subtests.

Differential weighting of the importance of separate tests is the second important question to consider. By intuition, there is nothing simpler than just summing up or averaging the separate subtests to construct the composite score. This method implicitly gives the same weight, which means using a "coefficient of importance" to multiply each of the tests, namely by the same weight $w = 1$:

$$z_c = \frac{1}{3}(1z_1 + 1z_2 + 1z_3).$$

If the subtests are not of the same importance, we can think about assigning them different weights w_1, w_2, w_3 to construct the *weighted average composite score* z_{cw}:

$$z_{cw} = \frac{1}{3}(w_1z_1 + w_2z_2 + w_3z_3).$$

Or, in general, for n subtests:

$$z_{cw} = \frac{1}{n}\sum w_jz_j.$$

For instance, we can take the relative importance of the first test, Z_1, as a base weight with $w_1 = 1$ as related to the other subtests. Then, say by expertise, we can decide that the second subtest, Z_2, is 3 times more important than the first test, and will be assigned the weight $w_2 = 3$. Finally, we find that the third subtest is just half as important as the first, so $w_3 = 0.5$. Then, for the patient in our example, the weighted average composite score would be

$$\begin{aligned} Z_{cw} &= (1/3)[w_1(1) + w_2(-0.5) + w_3(0.5)] \\ Z_{cw} &= (1/3)[1 \times 1 + 3 \times (-0.5) + 0.5 \times (0.5)]. \\ Z_{\mathrm{cw}} &= (1/3)[-0.25] \approx -0.08 \end{aligned}$$

Thus, considering the differential importance of the subtests, the overall score of the patient is not above average now, but it is a little bit below average:

$$Z_c = 0.33 \quad \text{vs.} \quad Z_{\mathrm{cw}} = -0.08.$$

2.3 Summary

The SD can be seen both as an attribute of the sample or population (the group's variability) and as an attribute of the test (the test's discriminability). Discriminability might be thought of as the ability of the measure to separate and distinguish individuals, or the ability to properly represent true individual differences on the test. A test with no or very little variance cannot distinguish individuals from each other. However, if there is substantial variability, but it is simply due to chance or error, it will again be unable to reliably distinguish between people with different true values on the

attribute. The sensitivity of a test can also be compromised by coarsening it into rough categories and losing information about true within-category differences. The variance of a single binary (0,1 coded) test is a direct function of its mean or proportion and is maximized when each of the two categories is equally probable (50/50). It shrinks and finally vanishes as the proportion approaches 0 or 1, and the test ultimately becomes useless at these limits.

Comparisons of test values for different people can be challenging if there is no objectively defined zero point of scale origin. Comparisons of values for different tests are problematic when the tests use different units of measurement or have different ranges of observed values. Where the values are all positive, "percents" of some maximum or other designated reference value are sometimes used to provide similar scales for different measures. Other approaches use measures based on norms, selected baseline values, and observed data ranges to provide a basis for establishing score comparability. One should also be clear about the connection between what one intends to measure, and the operations used for doing the measuring. With percents, it is important to locate a consistent origin or baseline for each measure, use a consistent principle for comparing the values to the baselines across measures, and take the possible ranges of the measures into account.

Measurement comparability is sometimes addressed through score standardization. Also known as the *Z*-score transformation, this is accomplished by subtracting the variable's mean and dividing by its SD. This transforms the original "raw" scores into *Z*-scores with a mean of zero and an SD of one. The resulting scores are consistently interpreted relative to their means and are scaled in the metric of SDs.

Z-scores having an approximately normal distribution can be mapped to relative percentiles of the normal density function and the cumulative normal probability function. The mean of this distribution is zero, 68% of the familiar bell curve distribution is between −1 and +1 SDs from the mean, and 95% of the distribution is between −2 and +2 SDs from the mean. Almost all of the distribution is between −3 and +3 SDs from the mean, or across a total range of 6 SDs.

Z-scores are a linear transformation of the original raw scores, while the associated normal distribution percentiles are nonlinearly related to the *Z*-scores. Changes, increments, or differences in *Z*-score values and normal percentiles thus have distinct meanings and interpretations. Also, differences near the mean or center of the distribution might not be comparable to differences of similar size near the distribution tails or extremes. Other standardized scores typically transform the *Z*-scores into convenient alternatives with means and SDs other than 0 and 1. For example, deviation *IQ* scores commonly use a mean of 100 and an SD of 15, while so-called *T*-scores have a mean of 50 and an SD of 10. Such transformations are typically used to give scores that are all positive with rounded integer values.

When the score distributions depart from normality, usual expectations about the relative position of scores will typically be somewhat altered or qualified. Skewed distributions and distributions with significant floor or ceiling effects can affect the interpretation of scores and score differences. Sometimes a non-normal distribution can be roughly "normalized" through a specific mathematical transformation such as taking the log or the square root of the original values. However, this can pose interpretational challenges or require back-transformation to represent in the original metric.

Having converted multiple scores to a mean-centered metric with SD units, we propose creating *Z*-score composites by calculating their mean. This can be appropriately applied where the components of the composite measure a single broader psychometrically justified attribute. Such additive composites of standardized scores implicitly allow compensation of low scores on one component by higher scores on another and weight each of the components equally when forming the composite. If differential component weights are desired, one can create a weighted average composite score with different weights for each component.

2.4 Study Questions

A. Discussion Questions

A Psychometric Look at the Variance and SD: "Discriminability" of an Item

Q1 Explain the two ways in which a variance or standard deviation (SD) can be interpreted in a *test calibration study* – as a characteristic of the sample, i.e. of the tested persons, or as a psychometric characteristic of the test itself. Also explain how this is parallel to the two ways a mean can be interpreted in a test calibration study, as stated in (Section §10 of Chapter 1 and Section §14 of the current chapter)

Q2 Why is it that a test with almost zero variance cannot effectively distinguish or "discriminate" between people, and cannot therefore provide a sufficient basis for establishing the validity of that test? Explain why we cannot also reason from this that a large variance or SD alone does not guarantee the validity of a test. (The second paragraph of Section §14)

Q3 Summarize Marion Richardson's logic in her 1936 Psychometrika paper that the content and scoring of a test in its initial calibration study should be manipulated by its developers to ensure that it has sufficient variance in the population of interest to establish "differential validity." (The third paragraph of Section §14)

Q4 How is the variance of a binary variable related to p, the proportion correct on that variable? In other words, what is the formula for the variance of a binary variable? (The first paragraph of Subsection 2.1.1.2)

Q5 Use the formula given in Q4 and the following simple binary dataset of 5 values to verify that, if you calculate a mean and a variance in the usual way on these values, you will obtain a mean equal to p (the proportion of "1" values) and a variance equal to pq. Data values for variable X: 1,1,1,0,0. (The first paragraph of Subsection 2.1.1.2)

Q6 Verify that a binary test item has its maximum variance, that is, its maximum "discriminability," when its difficulty level is $p = 0.5$. Using the formula from Q4, calculate the variance of test items having each of these four means (i.e. proportions): $p = 0.5$, $p = 0.4$, $p = 0.3$, $p = 0.2$, $p = 0.1$. (The second paragraph of Subsection 2.1.1.2)

§15 Using SD for Comparing the Incomparable: Percents vs. *Z*-Scores

Q7 What are "norms" and how they are necessary in test calibration studies? How does the word "norms" in psychometric usage relate to the concept of norms in everyday life as a "common external base"? What do norms such as means and variances have to do with comparing scales that seem incomparable? (The fourth paragraph of Section §15 and the entirety of Subsection 2.1.2.1)

Q8 "Percents are often assumed to be "unit free." Are Z-scores also unit free? How well do percents make valid comparisons of seemingly incomparable variables? How well does standardizing accomplish this purpose? Which is better? (The entirety of Subsection 2.1.2.1 and Subsection 2.1.2.2)

Q9 Conceptually and mathematically, how does the use of *Z*-scores, that is "standardizing" the variables, solve the problem of comparing variables that may seem incomparable? (All six paragraphs of Subsection 2.1.2.3)

§16 Density of Scores Along the Scale: The Normal Distribution of Frequency or Probability, the *N*-Law

Q10 From the text of this section, and Figures 2.2 through 2.5, explain what is meant by the "statistical theory of the *N*-law." What does it have to do with areas of the normal curve, and what are confidence intervals? (The first two paragraphs of Section §16)

Q11 Is the Gaussian curve ubiquitous in the natural world, a kind of law of nature? Where does it find its primary application in statistical theory and methods? (All three paragraphs of Subsection 2.1.3.1)

Q12 Explain the central limit theorem and its relevance for the use of the Gaussian curve in psychometric work. (Subsections **1.3.6.3** and **1.3.6.4** of Chapter 1)

Q13 Discuss the impact value of the *inverse normal cumulative distribution function* as shown in Figure 2.1 (in Subsection 2.1.3.1) for explaining and illustrating the relationship between the "bell-shaped curve" shown in part (a) of Figure 2.2 (in Section **§17**) and what it looks like when it is cumulated from left to right as shown in part (b) of that figure. (See all four paragraphs of Section §18)

§17 The Bell Curve or *Normal Curve Distribution Function p*(*Z*)

Q14 To understand how the pdf (*probability density function*) formula for the normal curve (shown in Footnote 2) works, follow and imitate Petr Blahuš's demonstration of his approximation formula using simple examples of $Z = 1.0, -1.0, 1.2, 2.0$, etc. as shown in the first three paragraphs of Section §17 and in Subsection 2.1.4.2

Q15 Summarize the discussion in the first full paragraph of Subsection 17b regarding being careful not to misinterpret the meaning of "probability density." In particular, note the statement that in looking for the probability of a particular *Z*-score location, the probability does not apply to "a point, but within a certain *interval*."

§18 The Ogive Curve or *Normal Curve Cumulative Distribution Function N*(*Z*)

Q16 (Birnbaum's 1960 formula for simulating the normal ogive, the logistic, is not only of historical interest. It also has theoretical interest in that it has its origin in IRT psychometric theory. (See Computational Guide 2 and Section **§52** of Chapter 5). Besides that, it also has pedagogical value as demonstrated in the following exercise: Do EQ1 in the Exercise Questions to get experience using the simple equation referred to as the logistic.

The logistic is of particular interest mathematically. Section §18 introduces it as a second approximation function but this time approximating the "ideally smooth" continuous normal ogive curve referred to as the normal curve cumulative distribution function, N(Z) which endows it with several mathematical advantages. As stated in the second sentence of that section, unlike the stepwise adding of discrete interval frequencies, the continuous ogive arises by integrating the density function p(z). The N(Z) density function is therefore recommended in the fourth paragraph of this section as being more exact than p(z). For an appreciation of the exactness of the logistic approximation see Figure 2.6 of Computational Guide 2.

§19 Percentile Position of a Score and the *N*-Law of Distribution

Q17 How were *Z*-scores used to compare the performance of the woman on the V-Fluency test to the performance of the man on the C-Praxis test? Once the *Z*-scores are calculated, there are three ways to find out the corresponding percentile (that is, the area below that *Z* value on the normal curve): (i) from a normal curve table, (ii) using the NORMSDIST function of Excel, and (iii) using the *logistic* equation from Section §18. Why does the text of Section §19 suggest you use the third (and most time-consuming) way? (§18 and §19)

§20 Are *Z*-Scores and Percentiles Fair?

Q18 Why does the author say "no" to the question of whether *Z*-scores are fair? Even though they are very useful for cross-comparisons between incomparable scales, they are not entirely fair for intra- as they are for interindividual comparisons along the same scale. What does this have to do with the relationship between *Z*-scores and percentiles? (§18 and §20)

§21 More Convenient Scores Derived from *Z*: The Deviation IQ, T-Score, and Stanine

Q19 For whom are these derived scores more convenient and why? (p. 21) Describe the "Deviation IQ" score, the T-score, and the Stanine score. Which are more useful and which are more problematic? (The four paragraphs of Section §21)

§22 What If the Scores Do Not Follow the *N*-Law? A Skewed Bell Curve

Q20 How does the T-scale violate the *N*-law of "the rule of 6 SD"*?* (The first paragraph of Section §22)

Q21 How do tasks that depend on persistence create a skewed or asymmetric curve? (The second paragraph of Section §22)

Q22 In the remaining four pages devoted to this section, the first author gives several examples of unusual characteristics of tests and the reasons for them. Did you find one or more that were particularly interesting? (Section §22 and Figure 27)

§23 Composite *Z*-Scores of Different Tests: Adding Apples and Oranges?

Q23 Summarize the discussion of the CERAD battery used in Alzheimer's disease research of how *Z*-scores can be used in combining subtests with very different numerical characteristics. (The four paragraphs of Section §23)

§24 Mutual Compensation, and Differential Importance of Subtests in the Composite

Q24 Explain "compensation" and the problems it can create for some ways of constructing an additive composite score. (The first three paragraphs of Section §24)

Q25 How can graphical test profiles help with some of the problems with compensatory additive composite scores? (The last three paragraphs of Section §24)

B. Exercise Questions

EQ1 As an exercise, solve the Iogistic formula (footnote 6) for one or more of the following *Z*-score values and then compare to the four place values in the normal curve table: $N(Z = 1.2)$, $N(Z = -1.2)$, $N(Z = 1.645)$, $N(Z = -1.645)$, $N(Z = 2.33)$, $N(Z = -2.33)$, etc. This will provide an insight into the nature of the Gaussian distribution and also help you remember critical ratios for 0.05 level, both one- and also two-tailed, and so forth.

References

Allen, E. J., Dechow, P. M., Pope, D. G., & Wu, G. (2017). Reference-dependent preferences: Evidence from marathon runners. *Management Science*, 63(6), 1657–1672.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstances for POMP. *Multivariate Behavioral Research*, 34(3), 315–346.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Quarterly Journal of Econometrica*, 47(2), 263–291.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika*, 1, 33–49.

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, 185, Issue 4157, 1124–1131.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference dependent model. *Quarterly Journal of Economics*, 106(4), 1039–1061.

Welsh, K. A., Butters, N., Mohs, R. C., Beekly, D., Edland, S., Fillenbaum, G., & Heyman, A. (1994). The consortium to establish a registry for Alzheimer's disease (CERAD). Part V. A normative study of the neuropsychological battery. *Neurology*, 44, 609–614.

3

Validity, Explanation, and Constructs

CHAPTER MENU

3.1 Validity and Validation, 99
- 3.1.1 (§25) The Validity Dialectic, 99
  - 3.1.1.1 Professional Standards, 102
  - 3.1.1.2 The Validity Concept and the *Validation Argument* Today, 103
  - 3.1.1.3 Measurement Instruments are an Inseparable Part of the Metrology Enterprise, 107
  - 3.1.1.4 Measurement in Any Field of Science or Technology, 108
  - 3.1.1.5 Summary: Validity Theory is Unsettled Historically and Theoretically and Needs Unification Again, 110

3.2 Terms and Tools in External Validity, 111
- 3.2.1 (§26) What is Validity? External, Internal, Concurrent, Predictive, and Practical Utility, 112
  - 3.2.1.1 Standardized Composite Scores, 112
  - 3.2.1.2 Criterion-Related Validity, 113
- 3.2.2 (§27) Is the Error of Estimating the Criterion Sufficiently Small for Some Test Uses?, 114
  - 3.2.2.1 Composite Validity, 115
  - 3.2.2.2 Variability and Validity, 115
- 3.2.3 (§28) Differential Weighting of Subtests: Composite Validity as Multiple Correlation, 115
  - 3.2.3.1 Weighted Composite Validity, 116
  - 3.2.3.2 Maximal Validity Via Multiple Correlation, 116
  - 3.2.3.3 Multiple Correlation to Shorten a Test, 117
- 3.2.4 (§29) A Criterion Score is Best Estimated by a Particular Composite, Multiple Regression, 118
  - 3.2.4.1 Comparing Several Methods for Constructing Composite Scores, 118
  - 3.2.4.2 Multiple Regression with Raw Scores, 119
  - 3.2.4.3 Standardized Multiple Regression, 119
  - 3.2.4.4 Regression Calibration, 120
  - 3.2.4.5 Error in Regression Prediction, 120
  - 3.2.4.6 Remarks on Interpreting the Magnitude of Multiple *R*, 120

3.3 Explanation Requires More Than Correlation, 121
- 3.3.1 (§30) *Scientific Explanation and Prediction*: Multiple Correlation and Regression as Its Statistical Metaphor (i), 122
  - 3.3.1.1 "Explained" Variance in Regression, 122
  - 3.3.1.2 Assessing Imprecision in Regression, 123
  - 3.3.1.3 Fisher's Information and Efficiency Measures, 125
- 3.3.2 (§31) Correlation vs. Causation: Their Respective Roles in Scientific Explanation, 125
  - 3.3.2.1 Experimental Manipulation and Passive Observation, 126
  - 3.3.2.2 The Counterfactual Condition and a Thought Experiment, 127
- 3.3.3 (§32) The "Cleaned Off" Incremental Validity of a Subtest: Partial Correlation, 127
  - 3.3.3.1 The Dilemma of Internal and External Validity, 127
  - 3.3.3.2 Partial and Spurious Correlation, 128
  - 3.3.3.3 Calculating Partial Correlation, 129

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

3.3.4 (§33) Scientific *Explanation of a Relationship* Between Attributes—Between Whole *Variables*: Partial Uncorrelatedness as Its *Statistical Metaphor (ii)*, 129
3.3.4.1 Partialing Via Sampling Stratification, 130
3.3.4.2 Partialing Via Statistical Control, 130
3.3.4.3 Local Uncorrelatedness and Falsifiability, 130
3.4 Internal Construct Validity and Latent Factors, 132
3.4.1 (§34) If the Explanatory Variable is Unobservable, *Construct* a Latent Factor (Construct Validity), 132
3.4.1.1 Latent Variables and Local Uncorrelatedness, 132
3.4.1.2 A Latent Common Factor Model, 134
3.4.2 (§35) What are the Behavioral "Constructs" Behind the Scenes? The Theoretical Concepts of Scientific Theory, 134
3.4.2.1 Theoretical Concept Formation and Latent Variable Modeling, 135
3.4.2.2 Latent Variable Modeling and the Principle of Partial-Local Uncorrelatedness, 135
3.4.3 (§36) Causal Effects vs. Independent/Dependent Variables: A Systems Theory View, 136
3.4.3.1 Causal Variables and Effect Variables, 137
3.4.3.2 Descriptive Associations and Causal Relationships, 138
3.4.3.3 Systems Flowchart vs. Path Diagram, 138
3.4.4 (§37) Theoretical Concepts as Active Subsystems Hidden in the S ⇒ R Black Box: Causal and Propensity Interpretation of Ability, 139
3.4.5 (§38) Theoretical Concepts as *Attributes* Under *Indirect Observation*, with Their Indirect Measurement Through the Manifest-latent Link, 142
3.4.5.1 Manifest and Latent Variables, 142
3.4.5.2 Reading Comprehension Ability as a Formalizable Theoretical Construct, 144
3.4.6 (§39) On Terminology 2: Mishmash Around Constructs, 146
3.4.6.1 Formalized (Mathematical) and Unformalized (Scientific) Variables, 147
3.5 Summary, 147
3.6 Study Questions, 148
References, 150

Why must I learn statistics to be a therapist?

After seeing the movie *Good Will Hunting* in high school, Maria was sure she wanted to be a psychotherapist. She wanted to pursue a profession that was interesting and rewarding—in her mind, psychotherapy was the perfect combination of those two things. To pursue her goal, Maria went to college and majored in psychology. She loved most of the classes, except for the ones that involved numbers. Maria had never liked math. And although she was generally a good student, she did not apply herself in her math courses. She was confused and disappointed when she saw statistics courses on the map of the psychology major.

"Why on earth do I have to learn statistics to do therapy," she would ask her professors. They usually answered by saying it was important for clinicians to be able to understand the scientific literature that informed research on psychotherapy, which can provide validated treatment options. Maria was pretty sure she could do that without having to calculate sums of squares by hand.

In her Junior year, Maria took the Graduate Record Examination (GRE), the test required for admission to graduate programs that trained therapists. Her GRE verbal score was good, in the 70th percentile. But her quantitative score was much lower, in the 20th percentile. As a result, most admission offices did not even review her full application. They merely looked at her scores and determined that she was not a good candidate for their program.

When she called to inquire about how she could improve her application for next year, she typically heard something like, "Your scores are too low. Retake the GRE and see if you can get a higher score, especially for the quantitative section." After hanging up, she would silently mutter to herself, "What is the connection between my math abilities and my clinical abilities?"

Maria's perturbed and frustrated question is a good one, and one that gets at the heart of the concept of validity, not just of tests, but of the nature of quality control issues in any domain, particularly the high-stakes work of clinical intervention. When Maria posed this question about her GRE experience, her concept of the variety of skills needed for effective psychotherapy was very limited.

This vignette now leads to an introductory discussion in Chapter Division 3.1, Validity and Validation, which is intended to broaden the concept of the domain of validity and the validation of tests and other forms of assessment to demonstrate that an understanding of statistics is needed. Petr Blahuš is good at connecting back to historical roots. The initial idea of a "validity coefficient" well over a century ago was the correlation between a test score and its criterion. In the next dozen or so pages, we will see how that concept has expanded and been transformed substantially over the intervening century.

Our discussion of Validity and Validation seeks to broaden the domain of validity concept and the validation of tests (and other forms of assessment) to demonstrate the need for an understanding of statistics. As Dr. Blahuš emphasizes, this has historical roots; the initial idea of a "validity coefficient over a century ago was the correlation between a test score and its criterion. The five subsections of Section §25 show how that concept has expanded and transformed in the intervening period. It gives a broad overview of validity theory to be revisited in other chapters when we deal with additional facets of this multifaceted concept.

3.1 Validity and Validation

Validity has been given a variety of definitions at different times and from different professional perspectives. We will elaborate on the currently dominant one in psychology and education, and show that it is unsettled. There exist strong counter-definitions from a variety of sources.

In current dictionaries, validity is defined generally as *the quality of being logically or factually sound*; secondarily, *the state of being legally or officially acceptable*. General dictionaries do not deal specifically with psychological and educational tests and measures. Professional standards, as we will see below, do so extensively.

3.1.1 (§25) The Validity Dialectic

In educational and psychological testing (including performance demonstrations), validation involves pragmatic concerns of an applied and even legal nature, and these concerns are currently being mixed, perhaps unwisely, with the scientific issues as validity theory evolves. In addition to testing, a related concept of validity applies to inferences from experimental and quasi-experimental research designs, a related application domain (Shadish, Cook, & Campbell, 2002). Good procedural designs are essential in both experimental method and also measurement validation, but the definitions of validity are not the same.

Early in the 20th century, most people in psychology and education thought of validity in psychological testing in simple terms:

> A test is valid if it measures what it purports to measure.

In writing these cogent words, Kelly (1927) clearly envisioned that *correlations* between test scores and criterion measures provided a kind of *validity index*. This was likely in the context of Karl Pearson's correlation coefficient, which he invented only a few years earlier. We cannot be

sure how else Kelley may have viewed other issues implied in these 12 words, words that are still refreshing and relevant for validity theory up to the present. First, the statement implies that *a test* can be valid or invalid. Second, test validity is about *measures*, that is, *measurement has occurred*. Interpretation and use had not yet become dominant issues. Third, also not central in Kelley's time, was the issue of whether or not differences in the actual property being measured *cause* differences on the scale interpreting the measurement outcomes. Correlation is not causation, but that was not a psychometric issue in 1927. Fourth, *purports* are like *claims*; meaning that someone (the test maker) has claimed to the buyer that there is something, some property, that the buyer might find valuable, and it is exactly that, not something else that the test measures. It is important to stand behind your claims, and when a test measures what it claims to measure it is deemed valid. That was the core of the first notion of validity.

Psychology and educational theory would endure operationalism, behavioralism, positivism, and other philosophical "isms" which made realism more difficult to pursue. Realism is now gaining traction at last, and causal approaches, rather than operational or empirical methods, can finally be considered. Realism assumes that a property (an attribute, a construct) actually exists, and that differences in its level of development in a person can enable that person to pass more items in an assessment and get a higher score. Also, the development of a new level of capability can cause an increase in relevant and adequate measures of that capability.

Samuel Messick is the validity theorist we will present first, in abbreviated form, in this brief introduction to validity. He has been characterized as a "constructive realist" in the book *Frontiers of Test Validity Theory* (Markus & Borsboom, 2013, pp. 10–11), which shows that 14 years after his death, he and his contributions to validity theory were still needing to be cited and explained. To the empiricists, he was stepping beyond the mark. Observations were considered to be the key to scientific advancement. Constructs were merely convenient organizing categories for theory development. But to Messick, the constructs had an independent existence. You have to "construct constructs" by finding the construct-relevant sources of task difficulty and by avoiding construct under-representation and construct irrelevance. But this careful design work is merely to enable the existence of constructs to be clearly manifested through persons' responses to items or performance tasks. This is also true of work in experimental psychology. Meticulous design is vital to avoid irrelevant and confounding factors and cleanly show the causal effects of the underlying actual nature of the independent variable.

Our senior author, Petr Blahuš was a scientist studying human movement—*kinanthropology*. He specialized in measurement—*kinanthropometry*. The "constructs" of human movement are real and visible. The performances of athletes and of aspiring athletes can be studied with video recordings at high speeds. The anatomy, physiology, and psychology of high athletic performance are a part of the story of a person's progressive attainment toward world championships. The strength and skilled movements, which are coachable when understood by great coaches, enable accomplished athletes and also athletes-in-training to score well in standardized athletic venues. Like Messick, Petr Blahuš did not refer to himself as any kind of realist but believed properly measured constructs would reveal their hypothesized effects through the construction of excellent and validated instruments.

Samuel Messick's unified theory of validity captured the prevailing view of the times. He described it as "Educators, with specialty in Educational Measurement," (Messick, 1989) in the 1989 Third Edition of the influential volume *Educational Measurement*, in the traditional chapter titled "Validity" (Messick, 1989). He also addressed the audience of "Psychologists specializing in Psychological Measurement" in the prestigious APA flagship journal *American Psychologist* in 1995 (Messick, 1995). This version also had a wide impact and lasting relevance. In these two

publications, he transformed the conceptual basis of validity theory in both areas. We illustrate his model of validity in Figure 3.1 based on a progressive matrix model introduced in these publications. His theory is complex. His definition of validity (Messick, 1995) reflects this complexity. The first sentence deals with the scientific essentials of validity of test scores/assessments.

> Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment (p. 741).

Since scores in the human sciences make the claim of being measures, we look first to the international definition of metrology to see how consistent this is with the more general and universal science of measurement. (See also Chapter 8.) The international definition of Metrology, the metatheory of measurement states:

> Metrology is the science of measurement, embracing both experimental and theoretical determinations[1] ...

Messick is true to this definition. He centers the validity of the measures produced by tests as an "overall evaluative judgment" of how well a combination of empirical evidence and theory supports this judgment. Without mentioning it himself, he is consistent with the science of metrology here. It is peculiar that his definition does not include any mention of excellent measurement, but he clearly championed it elsewhere. He was facing the big task of encouraging the field to pay attention to consequences and values, and may have assumed measurement needed no defense. While consideration of human values is avoided by hard scientists, Messick involves values in the same overall evaluation along with hard evidence. One of the controversies he faced has been over his (and other validity theorists') insistence that values and consequences are an inseparable part of the

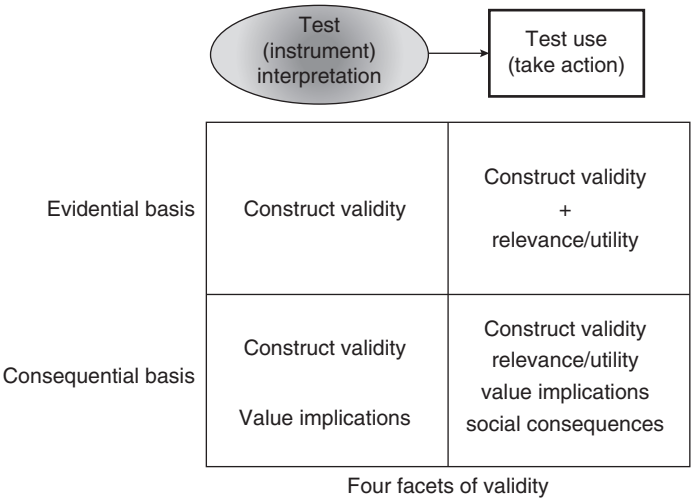


Figure 3.1 Facets of validity as a progressive matrix.

1 The Metrology definition continues: "at any level of uncertainty, in any field of science and technology."

overall validity concept. Figure 3.1 represents the unified validity theory. Note the two rows, Evidential Basis and Consequential Basis, and inside the cells in the consequences row, how consequences give rise to values.

Messick's definition of validity introduces how he narrows the scope of validity—to test interpretations and test use. Test users make decisions and take actions based on individual test taker's scores. This includes forms of assessment other than tests such as rating scales for observing performances. This model directs us to seek evidence both from the item or task data that produced the test scores and other validation evidence, including the theory behind each construct. The two columns of Figure 3.1 represent interpretations and test-use actions, crossing with the rows of evidence and consequences. He further limits the scope of this otherwise enlarged and encompassing theory of validity by leaving out the test itself, that is, the testing instrument, with the next sentence following the definition quoted above:

> Validity is not a property of the test or assessment as such but rather of the meaning of the test scores.

Some have objected to the omission of the measuring instrument itself from the definition of validity, and in the spirit of dialectic, perhaps rightfully so.

It is called a progressive matrix model. Using the test results to take some action must follow interpretation, so the matrix progresses from left to right. It also progresses from top to bottom, because evidence of construct validity and relevance/utility must be the foundation for value implications and social consequences. Construct validity is the first entry in each cell, serving appropriately under the meanings of row and column headings. Like his predecessor who authored the preceding chapter on Validity in *Educational Measurement*, Lee J. Cronbach, Messick agreed that the day of different kinds of validity ("content validity," "criterion validity," etc.) was gone. They are all aspects of construct validity need different kinds of validation evidence from both internal and external sources.

Messick (1995) listed six aspects of construct validity: (i) Test content, (ii) Substantive process, (iii) Score structure, (iv) Generalizability, (v) External relationships (it includes criterion measures), and (vi) Testing consequences. You can see that so-called "content validity" became content-related aspects of construct validity, and "criterion validity" became part of external relationships. Construct validity gobbled them both up, and rightly so. It is useful to note that the first three are *internal aspects* of construct validity. These deal with analysis and data from the test itself, while four through six are *external aspects*. They require external evidence from additional samples of people or from additional measures not part of the testing instrument itself. These six, and in addition, relevance/utility, and value implications offer a large array of aspects of validity for which validation evidence can and should be sought, according to the Messick approach.

3.1.1.1 Professional Standards

Professional standards represent a consensus-finding process for best practices across fields for psychometricians and others using tests and measures. Their progression over time captures consensus in educational and psychological testing and reveals the extent to which newer concepts become accepted and formalized in standards.

In his 1995 article, Messick put validity in context with other concerns about test development processes and use. He spoke of validity, reliability, comparability, and fairness. These, he said "*are not just measurement principles, they are social values that have meaning and force outside of measurement whenever evaluative judgments and decisions are made*." They are all measurement

principles—and more. These four concerns are covered in a variety of ways throughout the major sections in the standards for testing, which are jointly published by three professional organizations: AERA, APA, and NCME.[2] The latest are the 2014 *Standards for Educational and Psychological Testing*, which is the 5th version, the previous two being in 1999 and 1985.

In order to develop a new *Standards*, a writing/editing team is set up with noted professionals from one or more of the three organizations as chairs and co-chairs of the committee. They work with a large set of consultants who read drafts and give hundreds of incisive, referenced comments. These are discussed by the editing committee and the draft is modified, then published as the newest version of the *Standards*.

Looking specifically at the evolution of the Validity concept, the 1985 *Standards* shows that the ideas Messick incorporated into his Unified Validity Theory were understood and accepted by the writing committee and their many consultants. His ideas were part of the zeitgeist of the time, and he articulated and shaped them well. The 1985 *Standards* was an important bridge. It was the first to dispense with the idea of different types of validity (construct, content, criterion), but the first chapter (which is traditionally focused on validity) still retained some vestiges from the old view of "validities." It was still organized under headings such as Construct-related Evidence, Content-related Evidence, and Criterion-related Evidence.

A brief summary of the next two *Standards* shows they continued with the incorporation of separate aspects of construct validity, not "kinds" of validity. The 1999 Standards also added consequences. *Consequences* were even more strongly differentiated in 2014, with an emphasis on unexpected consequences. As for the six aspects of construct validity, they continued through 2014 but only under five headings (generalizability was grouped under external validation).

Since it has been so long since the last standards were published, a website was set up for free sharing: https://www.testingstandards.net/open-access-files.html. It makes available pdf copies of the 1999 and 2014 standards in English and Spanish, and an eBook for the 2014 standards. Students who are serious about a career in psychometrics are encouraged to download these open access booklets.

The *Standards* guide best practices. They are not laws and have no mechanism for enforcement, but good lawyers for plaintiffs well understand them and the prevailing civil law, and use them against test developers or test users who took actions they argue were hurtful to their plaintiffs. The plaintiffs are usually individuals or groups who took the test and objected to a use made of it that cost them an admission, a job opportunity, etc. These professional standards provide *official standards for approval*, but they do have *legal implications*.

3.1.1.2 The Validity Concept and the *Validation Argument* Today

Messick's successor in writing the next Validity Chapter in *Educational Measurement* was Michael Kane (2006). But like Cronbach (1971) in the *Educational Measurement* volume earlier, Kane titled it Validation rather than Validity, giving it a more applied and pragmatic slant, while trying to keep in balance a scientific approach. Since 2009, Michael Kane has held one of the five prestigious funded chairs at ETS—The Samuel Messick Chair in Validity Studies. He is seen as the champion and spokesman for validity argument—obtaining validation evidence that argues for the veracity of claims that instrument developers make for the purposes and uses of their testing products.

2 American Educational Research Association, American Psychological Association, National Council for Measurement in Education.

Kane was not the first to speak of *validity argument*, but he has long been a most influential advocate and innovator in it (see Kane, 1992, 2006, 2009, 2013a). Lee Cronbach proposed *validity argument* in a keynote address to an audience of validity experts at a conference held at Educational Testing Service on May 28–29, 1986 (proceedings published in Braun & Wainer, Eds., 1988). In the keynote address, Cronbach stated: "...I propose here to extend to all testing the lessons from program evaluation ... I invite you to think of validity argument rather than validation research."

In Michael Kane's (1992) paper, he established an early leadership as advocate and definer of *validity argument*—a leadership which he has maintained. This reached a high point in 2013 when he took the lead in a full issue of the *Journal of Educational Measurement* on validity. Kane's (2013a) lead-off article, *Validating the Interpretations and Uses of Test Scores* was followed by five strongly or mildly salutary responses and one strong rebuke by Borsboom and Markus (2013). Their article showed that evidence and argument can be persuasive even in the absence of *truth*. Is there no place in validity theory for evidence that the thing being measured truly exists? Markus and Borsboom used the once popular phlogiston theory of fire to illustrate how scientists of the day found evidence to argue for a non-existent substance.

Steven Sireci led the cheerleaders for Kane and *validity argument*, with these words:

> I have long been a fan of Dr. Kane's writing on validity (e.g. Kane, 1992, 2006) and consider him to be one of the greatest validity theorists of our time. The "argument-based approach" to validity he articulated in Kane (1992) was essentially incorporated into the most recent version (1999) of the *Standards for Educational and Psychological Testing* ...

However, Sireci, an astute scientist and scholar, could see that requiring a new argument for every purpose and every interpretation is at the very least, a costly and tedious requirement. He offered a way to shorten Kane's approach to three steps. Another writer, Paul Newton (2013), also proposed a way to shorten it. However, in his final article in this journal issue (Kane, 2013b), Kane disagreed with these two that his methods should be shortened, and dismissed Borsboom and Markus's argument because he did not accept their proposed conception of validity.

When the *Standards for Educational and Psychological Testing* came out the following year, 2014, validity concept and validation argument were strong and central. How did Messick's complex view of construct validity fare? There was no strong and articulate champion to clearly explain the progressive model nor how to use it in practice. Is there a future for the *validity argument*? Will Messick's progressive model centered on different modes of construct validity prevail?

A 2009 book titled *The Concept of Validity* (Lissitz, 2009) came at an important juncture to address this question. When we began work on this book, it was included among the references that we inherited with the original manuscript from Petr Blahuš, which highlighted its impact on us as successor authors. The book truly provides strong objections and challenges to many aspects of Messick's received model, as well as many defenses. This book was organized and edited by Robert Lissitz from presentations made at a conference in 2008. In the introduction to this volume, Lissitz stated: "Sadly, in my opinion, the unified field of validity as articulated by Messick (1989) is not just historically unsettled at a theoretical level, but it offers little in the way of useable advice to people working on test construction ..." He also mentioned how difficult construct theory was to understand, and cited several other professors, difficulties in explaining to students or to lay people the

integrated view centered on construct validity. As Lissitz pointed out, it does not pass Einstein's "explain it to your grandmother" test.

Borsboom and four European co-authors (Borsboom et al., 2009) gave their chapter the ominous, or encouraging, title (depending upon one's point of view) of "The end of construct validity." They objected that after Messick and Kane, a test could no longer be considered valid if evidence merely indicated that it measured what it purported to measure. In 2004, Denny Borsboom and his colleagues put measurement back at the center by restating the definition of validity in this way:

> A test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure. (Borsboom, Mellenbergh, & Van Heerden et al., 2004, p. 1)

They ignore the mainstream view of Messick, Kane, and Sirici that the testing instrument itself has no property of validity, and do not mention that interpretations and actions should be included as part of the validity concept. They are not alone in this view but are quite outside the mainstream.

Borsboom's colleague and graduate school mentor, frames it in this way, "Societal aspects of testing have to be excluded from psychometrics because ethical and political issues cannot be settled by scientific methods" Mellenbergh (2011, p. 457), (p. 445). In a similar vein and on the same page of this book, Mellenbergh gives a simplified summary of the two positions:

> In the field of psychometrics, opinions on the consequences of testing diverge. Some authors consign the consequences of using tests to the concept of validity (see, among others, ... Kane (2006), Messick (1989), etc.), whereas others think that consequences have to be excluded from the validity concept (see for example, Borsboom (2005), Borsboom, et al. (2004), etc. However, both groups agree that societal aspects of testing are extremely important.

Markus & Borsboom (2013) published a thoughtful and persuasive book *Frontiers of Test Validity Theory*. This book does not have any of the strident tones about "the end of construct validity" nor even how easy it will be to have a causal model of validity. These two authors succeed in showing that we still have much to learn about the three broad domains of measurement, causation, and meaning in test validity theory and research. They do, however, also point out that despite Messick's focus on scientifically verifiable constructs, he switched the definition of validity away from *scientific measurement* by focusing on interpretations and actions. Messick emphasized the evaluation of evidence, an *epistemological* concern, rather than the more fundamental *ontological* concern of *factual reality*, that is, whether the thing being measured even exists. In so doing, Messick paved the way for Kane's success in making *Validation Argument* the main focus of Validity Theory.

Mislevy (2009), in his chapter of the Lissitz book, is comfortable with and often quotes Messick. Moreover, he has long used Toulmin's (1958/2003) work on supporting claims with evidence. Toulmin comes from the perspective of legal reasoning to find evidence both in favor of the claim and to rebut alternative reasons. Toulmin's work is fundamental to *Validation Argument*, and Mislevy has long been a strong voice in bringing forward the Toulmin work on providing evidence to support claims. Do not mistake, Toulmin's work is not restricted to legal claims just because that was his background and domain of operation. It can be any kind of claim. Toulmin's work also comfortably leads to evidence for scientific claims, such as whether a test measures what it purports to measure, and also explanations of what is causing changes in the test scores.

Mislevy raises fundamental issues because he does not defend "constructs" at all, even in his discussion of Borsboom et al.'s "end of construct validity" chapter. Instead, he invokes specifically two additional sources of theory, socio-cognitive aspects of the test taker and model-based reasoning, with indirect mention of *Design* as a source and method in model-theory development. In so doing he identifies himself as a constructivist-realist in the use of models. Mislevy has contributed a more recent and more thoroughgoing treatment of design in his book *Sociocognitive Foundations of Educational Measurement* (Mislevy, 2018). However, in this earlier 2009 chapter, he was starting more simply. Here he begins with cognitive theory using the work of Snow and Lohman (1989), then goes to a more complex sociocognitive level as he brings in the linguistic and cultural background of each individual who interacts with the assessment instrument.

As Mislevy explains, Snow and Lohman's paper on the implications of cognitive psychology for measurement offers an alternative to the notion of a trait expressed by Messick and many others, that is "*a relatively stable characteristic of a person…consistently manifested despite considerable variation in the range of situations …*" Instead, the view from cognitive psychology is that "*test performances comprise complex assemblies of component information-processing actions that are adapted to task requirements* ***during performance***" (emphasis added). Snow and Lohman warn that both scientists and the public have been misled into believing that fundamental fixed entities are being measured. Mislevy is not one of the misled. Mislevy (2009, p. 101) declares decisively that "*Person parameters such as θ and η in psychometric models reside in the mind of the analyst rather than the mind of the subject*." He also states in the first sentence of his abstract to this chapter that "*between-persons variables in trait-based educational assessment are absurd oversimplifications. Yet, for a wide range of applications, they work*" (p. 86). In this chapter, Mislevy also shows why they *are* absurd oversimplifications, using not only the Snow and Lohman paper but going deeper with the sociocognitive perspective, which demands four-way relationships among the model, the system, the user, and the purpose. It is at this level of detail that what is really happening can be seen, and it is deeper than is usually given in the psychometric models used to summarize the data, or in the traits or other constructs based on between-persons variables.

What Mislevy offers is an introductory lesson in model-based reasoning, a more powerful scientific approach than constructs based on between-persons variables. He does not promote his own evidence-centered design (ECD) concepts directly, but makes the power of models clear to those who understand the first two layers of ECD, that is, domain analysis followed by the synthesis of models. Model synthesis includes models of the student, not just θ or η but also models of the tasks and evidence models for key claims. He explains domain-specific model-based reasoning in a way that shows it to be more self-critical and more connected to empirical data and task variables than vague general "constructs." He shows other benefits in his discussion of good instrument design and improvement processes. He seems to provide what Lissitz said was missing in the unified validity model—understandability and advice for the development of valid measurement instruments. He does vindicate psychometric models at those times when they work well enough to be useful. But on the introductory page (p. 4), he quotes Box and Draper (1987) that "*all models are wrong; the practical question is how wrong do they have to be to not be useful*."

We might add that the ECD contribution of Mislevy and his colleagues also completes a larger progressive matrix for Figure 3.1. Going back to that figure, we now add three ellipses, and their attending process arrows, to the left of the current starting point of "test instrument interpretation," the shaded ellipse.

Interpretation of the testing instrument must be preceded by the creation of the instrument in three distinct stages: (i) Designing the instrument, (ii) Creating the instrument from that design, and (iii) Scaling and then administering the instrument. It is only after test administration that the two columns of the unified model are mentioned, as (iv) Test interpretation and (v) Test use (shown shaded and in shapes like Figure 3.1).

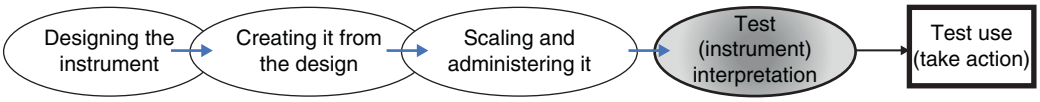


Figure 3.2 Expansion of the columns of Figure 3.1 to accommodate Mislevy modeling.

3.1.1.3 Measurement Instruments are an Inseparable Part of the Metrology Enterprise

The wording of both Messick and Kane is very careful. Messick (1995): "Validity is not a property of the test or assessment as such ...," and Kane (2013a): "it is the proposed score interpretations and uses that are validated and not the test or the test scores."[3] Sireci (2009) also notes that [Validity] "is *not* an inherent property of a test." The word *test* is ambiguous. Some readers would consider the word to designate the testing instrument, as in the old "A test is valid if ...," but the high context specificity of the *Validation Argument* idea is that these writers are referring to a *particular testing occasion* where interpretations for particular purposes are made. That is the context for the terms "test and test score." They wisely for their purposes, but ambiguously, do not use the word instrument in the same breath with validity.

The word *instrument*, however, pervades international metrology. It is only through special scientific instruments that most physical and chemical properties can be measured at all. Historical progress in better optical instruments (and shorter waves than light) has repeatedly revolutionized the study of the very large and very small. Other instruments have done it for the very hot and very cold. The Large Hadron Collider enables physicists to "observe" the invisible collisions of proton streams with particles of interest. Not just science, but engineering and technology jump to a new level of effectiveness when better, sometimes larger and more powerful, and other times smaller and more portable yet powerful instruments are developed through applied metrology. Instruments are so fundamental that they are found in definitions in the International Vocabulary of Metrology (VIM). The instruments can also have built-in calibration and traceability evidence for the scientific definitions. Systeme international (SI) instruments must meet international standards with evidence for claiming that they measure to a certain level of uncertainty, so they clearly pertain to the measurement issues of comparability and reliability. Their implementation plans and contents should meet professional standards for fairness. Legal metrology picks up on this whenever laws apply to measurements. It is inconceivable that validity has nothing to do with instrument integrity—unless our concern is only with whether a proven instrument is valid for a particular testing occasion and a particular use.

The word *instrument* also pervades education and psychology. A viable test instrument in these domains is also the result of a careful design and development process, along with an implementation design process to guide its use and the interpretation of its scores. The *Standards* specify many details of test development and documentation that contribute to validation evidence during development, and many of these are related to the components of the test instrument. The relevance of instrument integrity and quality to validity is not found only in the validity chapter of *Standards* but in other chapters dealing with the reliability, comparability, and fairness issues mentioned by Messick.

Designing, developing, and putting a new testing instrument or battery into use meets the historical understanding of where validity starts. Angoff (1988) in reviewing the history of the validity concept reminds us of this reality by quoting Anastasi (1986) "Validity is ... built into the test from

3 Perhaps the conflicting views of the two camps of validity theorists is at least partly a function of two or more meanings of "valid." Even a test that has been proven to be ontologically valid can be applied in an inappropriate context that renders the score interpretation invalid.

the onset rather than being limited to the last states of the test development, as in traditional criterion-related validation." The same may be said of the validation evidence assembled in a *validity argument* for a specific claim. How are the items developed? What are the test's purposes and how are they implemented, executed, and documented? How is scoring set up? What are the cut scores, and how they are obtained? Are norms developed, and how they are developed? Do the items function differently for different groups? How are users taught to use and administer the test properly? It is not difficult to see why the foundations of validity, not to mention reliability, comparability, and fairness, must be built into the instrument and its documentation package from the start. It is not difficult to see why the progression that ends with *interpret* and *use (take action)* was "designed on purpose" from many of these earlier constructions, with the many empirical tests and revisions that should have taken place.

Now in addition to the rather general standards, we have multiple volumes explaining the applications of *evidence-centered design of assessments* (ECD) and *Principled Assessment Design for Inquiry* (PADI) in a variety of domains and situations, available without cost courtesy of government funding. (See chapter references, available from SRI International[4]). It enables designers of assessment instruments to investigate each domain of interest to institutions and individuals, organized by the first layer of ECD—domain analysis. Then in the second layer of design, domain models, is guidance on how to understand the domain processes by model synthesis as well—creating student models, task models, evidence models, and others. Models and Theory must be there for validity to be an evaluation of both evidence and theory. It is not hard to see why Mislevy (2009) chose not to dwell on constructs—such a vague term outside of the unanalyzed domain of interest. Models and model-based reasoning replace *so-called constructs* with domain- and function-specific models. After a domain analysis, the domain-specific and intended audience-specific boundaries have been determined. These boundaries and content varieties enable the development of models for students, tasks, evidence, and other matters. Also, the implementation requirements are planned during the design layers. The scaling decisions are made and the instrument is calibrated to ensure it fits a comparable and interpretable measurement scale.

Good validation arguments for the claims of the instrument makers are essential. Kane has emphasized the pragmatic, but he is also a good scientist. "Pragmatism" does not sound like the *science of metrology*, which is where it all begins. A pragmatic argument clearly spans the broader field of metrology, *scientific*, *applied*, and *legal*. Its center in *Validation Argument*, however, appears to be somewhere within the pragmatic *applied* and substantially overlapping the *legal*. Now we turn to Metrology as the science of measurement.

3.1.1.4 Measurement in Any Field of Science or Technology

The new book *Measurement Across the Sciences* (Mari, Wilson, & Maul, 2021) is a collaboration of an engineer-metrologist, Luca Mari, who has participated in the editing of the current Vocabulary Internationale of Metrology (VIM 3), and two versatile and experienced psychometricians, Mark Wilson and Andrew Maul. It is based on 10 years of collaboration and a number of intervening articles. It has the bold subtitle *Developing a Shared Concept System for Measurement*. This new book places international metrology in a central position and resonates with the definition of international metrology cited above and in chapter 8.

Metrology is the science of measurement, in any field of science and technology, at any level of uncertainty. This 2021 book is a harbinger and door-opener inviting the Psychological/Educational Measurement communities and all human sciences to join the world's scientific-metrology community, with all its history, standards, definitions, methods, and great tradition of vanquishing multiple incompatible measures and creating social and economic capital.

4 Stanford Research Institute became SRI International in 1977.

To present and justify a definition of the necessary foundations of measurement applicable to all sciences, they explain in their Chapter 2 that measurement has four defining characteristics. They started with VIM 3 definitions, which assume ratio scales in almost all the units accepted as part of the SI, international system of units. Then they created a new definition based on four defining characteristics that they hope will define measurement for all sciences. The four are italicized below. Our comments follow the italicized name of each progressive aspect of the definition of measurement.

1) *An empirical process*—it must be a *standardized empirical process* to be measured.
2) *Designed on purpose*—it is a designed process, with intentional purposes. In all sciences, including human sciences and applied technologies, professional groups of people share common purposes in each of many *domains* of interest and activity. These domains encompass *objects* that may be physical, biological, or human. Many of these objects have *properties* of value in each domain's human activities. Many of these properties can be measured. The raw measure of the property must be converted to *values* on an informational scale that has a defined unit and a meaningful scale.
3) *Whose input is a property of an object*—measurable properties of human objects must have empirical consequences such that when persons apply their senses, mind, and hands or voice to a task and make a response, the property is *transduced* to an intermediate value within the instrument. "Transduction" is a key term and concept in this definition.
4) *Whose output is information in the form of values of that property*—Mari et al. use a black-box model to explain the parts of their cross-sciences definition of measurement. The black box explicitly contains an intentionally designed instrument that *transduces* something empirical about the property to an intermediate value on the instrument's internal scale, as explained in 3. The instrument or its user matches the reading to a number on a local scale. Then the instrument applies *calibration mathematics* to compute a value on a clearly informative *public* scale. This scale provides meaning for interpreting different levels—values—of the property. It should provide quick and easy interpretation usable in what the authors call *the informational world*. To measure the property, it must be a scalable property. Objects with their properties exist in the world we look at as real, while public scales and the meaning of particular values belong to what we can call the informational world. The intentionally designed instrument has a lot of work to do, as we see, when we peer inside the black box and find inside it explainable workmanship and function.

The metrology authors are starting with existing definitions in VIM3. They are trying to create and offer a version of VIM that will embrace the nonratio scales of the human sciences, perhaps even our often less-than-perfect interval scales. International units now include the Mohs scale of increasing hardness, which is an ordinal scale—a hopeful precedent.

The fundamental problem the authors must overcome is that the current definitions in VIM speak of "quantities" which are ratio scales. The word "quantity" infuses the definitions of the VIM, and thus the definitions of the SI base and derived units (see chapter 8). The word "value" is used in VIM to refer to values on a (usually quantitative) scale. But Mari, Wilson, & Maul are using the word "value" in place of "quantity" to make it possible for other forms of scales for use in "all sciences and technologies." The VIM already acknowledges ordinal scales and is even proposing nominal scales for categories. The Mari et al. book was written at a time when changes such as these authors propose could actually take place.

The fortuitous double meaning of the word "values" is delightfully true and encompassing—(meaning 1) a numerical value on a meaningful, easily interpreted scale, and (meaning 2) something of great social and economic value in a particular domain of objects important to human

users of the instrument that produces the informative *values*. It is important to them—very valuable—because it has important consequences for bettering or managing their lives. One cannot understand domains of interest without understanding the consequences and values they hold for the people interested in them. The *Values & Consequences* part of Messick's definition must continue in the Standards for Psychological and Educational Testing, but going beyond the traditional concern about negative consequences and unfairness to the positive value of human attributes that are teachable and learnable. These must be a significant part of the domain analysis for each instrument.

What is the cross-science definition of measurement? And what is the definition of validity? In the new book, the new definition of measurement proposed to apply *across the sciences* is:

Measurement *is an empirical and informational process designed on purpose, whose input is a property of an object and that produces justified information in the form of values of that property*. (This is a generalization of the VIM definition in 2.1 of VIM 3).

And what is the definition of validity? In VIM, *validity* is not a defined term. The closest is "instrument accuracy." Mari et al. may soon discover that *interpretations* of the values produced by the instrument, *and evidence of claims for uses for particular purposes*, are a part of the informational world's use of any value produced by any instrument. The contributions are of inestimable value. Yet, they are missing the full concept of validity needed in their *informational world*.

In his recent volume, *Sociocognitive Foundations of Educational Measurement*,[5] Mislevy (2018) gives us a synthesis that brings measurement back into the definition of validity, and also keeps interpretation and use. In the Lissitz book, Mislevy had agreed with Borsboom rather than Messick that validity was not a judgment. He would correct that by replacing *judgment* with *the degree to which evidence and empirical rationales* justify or support the reasoning. After 10 years of preparing his sociocognitive synthesis, he would modify Messick's definition as follows.

> A test is valid for a given interpretation or use to the degree to which empirical evidence and theoretical rationales support reasoning as if "it measures what it purports to measure."
>
> (Mislevy, 2018, p. 202)

He offers this definition as "*a constructivist-realist variant that reflects the "as if" nature of the use of measurement models*." He has not wavered from his 2009 socio-cognitive view that there is no real unified and stable property in what is being inferred from between-subjects data but continues to acknowledge that our models can often be good enough if our model-based reasoning is sound, sensitive to good sources of empirical evidence, and self-critical.

In line with the questions posed by scholars like Lissitz, Markus, and Borsboom, and now Mari, Wilson, & Maul, the future of the concept of validity and the practice of validation research, whether it is called *validity argument* or not, is open to better theoretical formulations.

3.1.1.5 Summary: Validity Theory is Unsettled Historically and Theoretically and Needs Unification Again

We use some of Lissitz's (2009) words in this heading, and back up our summary below with the material and references given above. At this point, we have this to say in summary.

1) We regard it as problematical that out of the rightly separate triad of Scientific, Applied, and Legal Metrology, the current *Validation Argument* position is so un-selfconscious of the extent to which Legal Metrology and Applied Metrology have dominated their position. *Validity*

5 Called "a conceptual and methodological tour de force for the field of Educational Measurement" by James Pellegrino.

argument can correct this given that its centering and evidence collection can broaden to include the scientific modeling of the test takers, and how the instruments empirically make contact with the varying levels in the models of student progression. They can include models of the instrument and its implementation to reduce threats to validity and can draw upon both instrument design and implementation design to provide the necessary evidence for key claims.

2) Borsboom and others make the point that seems will eventually prevail. Historically and rationally, a *measurement* instrument should *measure* what it claims to measure. That should not be forgotten again and should never change. How far his ontological and causal definition will go remains to be seen. Mislevy (2018, p. 202) has offered a new definition cited above which bridges the chasm between the Messick version and the ontologically sound earlier definitions of validity. We favor the Mislevy synthesis at this time.
3) Kane's case that claims should be validated properly, and that purposes and interpretations are central will also continue to hold their importance. Sociocognitive perspectives will only deepen that importance. But validity theory must give room again for valid, accurate measurement instruments. Science progresses that way.
4) Kane also makes the point that you can have more narrow purposes and interpretations, and have simpler definitions of validity. He is correct, but it should not be an either-or between the Borsboom and Kane 2009 positions and will not be. Borsboom's position has grown more circumspect (Markus & Borsboom, 2013). Kane asserts that stronger claims (e.g. trait or construct theories) are usually more important but also more difficult to validate. We would agree, but add that they are much more important if the theoretical model or progression of models is to become recognized as a scientific reality to improve interpretations and guidance to learners.

3.2 Terms and Tools in External Validity

The four sections of this second division of the chapter, Sections §26–§29, provide an account of the first concept of validity in what is now the field of Educational and Psychological Measurement: *test-criterion correlations*. It is perhaps the most interesting aspect of external validation, but one that has never explained how to obtain a true criterion measure. Blistering critiques of using correlation as a validation coefficient are found (e.g. Borsboom et al., 2004). Prof. Blahuš is just getting started, however, and in Section §31 and also in the third part of this chapter (3.3 Internet Construct Validity and Latent Factors), he will discuss causation vs. correlation. By the end, he will have helped us understand why he uses quotes around "constructs" and uses "so called 'construct validity'" on the often long path from poorly conceived and poorly instrumented "constructs" to real scientific variables. We believe he might be quite pleased with the approach of Mislevy and others in PADI for developing domain specific models for both student progressions and task difficulty—no need to talk about constructs.

The last section in this second part of the chapter, Section §29, wraps up Petr's methodological elaboration on the various ways of thinking about the relationship between a criterion score and one or more predictors. It also completes a discussion of multiple regression, and what it means and does not mean. Appendices 4 and 5 are indispensable guides with tools to give each student a better mathematical grasp of multiple regression, and to use software to investigate data sets using it. This lays the foundation for a discussion of deeper and more powerful ways to look at data and relationships among variables—latent common factor (LCM) models.

In these three part of the chapter, the senior author mentions some other terms ending in "validity," including absolute/differential validity and composite validity. These are **not** additional

types of validity. They are terms tied to tools and provide a way to talk about different analytic treatments that have been applied to modify a correlation-based validity coefficient. More is found to complete the topic of Validity and Validation methods in Chapters 4 and 5.

3.2.1 (§26) What is Validity? External, Internal, Concurrent, Predictive, and Practical Utility

Criterion-related validity is a place to venture into the broad topic of Validity. It is an external aspect of validity—relationships to measure external to the scores produced using a particular instrument. External aspects of validity include not just relationships with other measures, but also external influences, like cheating, poor conditions during a test, or different effects of the test and certain of its items on people with different genetic and cultural backgrounds. External influences can represent potential threats to validity—evidence of unfairness and accompanying invalidity. By contrast, relationships to strong criterion variables, both predictive (future external measure) and concurrent (external measure administered at about the same time), provide evidence in favor of validity. This external evidence can help establish an argument for the *external* aspects of validity—generally over time, not all at once. Internal aspects of validity center around the actual content, structure, and questions in the test and how they match the actual substantive thinking processes (or actions) that a person uses to do well in life. Think about those invisible abilities a person uses situationally both to do well in life challenges, and at another time to get a higher score on the test. They should be the same. The measurement practitioners need to *construct* some idea, some model, of what those invisible talents are—a construct. The test and any good criteria must tap into the same abilities, knowledge, skill, personality trait, or disposition for the test to be a valid way to assess those characteristics of persons. Constructing such a model of those invisible capabilities of the person enables instrument designers to develop tasks/questions for the test instrument.

Later in this and other chapters, using the basic knowledge of regression and correlation, we will go deeper into the topics of validity, reliability, and further psychometric aspects of the diagnostic quality of a test. For now, we will just touch on one aspect of external validity, the correlation of the score with a criterion measured separately from the test.

Quite often, we wish to assess the quality of a test by relating or comparing it to another well-established diagnostic method—sometimes called a "gold standard"—to verify whether the test scores agree with it reasonably well. Using a previous example,[6] we may wish to verify an association between a composite score from the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) test battery and the clinical diagnostics from the Mini Mental State Exam (MMSE). Sometimes, we would like to make a prediction of the suitability of an applicant for a certain kind of job, predict how well a test is able to indicate the future progression of a psychological disorder, etc.

3.2.1.1 Standardized Composite Scores

For brevity, let us consider our previously discussed three selected subtests of CERAD as a composite, use the z-scores of each, and denote it as *a composite of z-standardized score*s. We will refer to it simply as z, which also includes unweighted (or equally weighted), z_c, and weighted, z_{cw}, composites while the MMSE clinical result is represented by y. This suggests as a natural and simple possibility, the use of the correlation coefficient r_{zy} as a measure of the coincidence between the two diagnostics. The relationship between a CERAD composite score z and the MMSE diagnostics y might also be of interest for assessing the "practical utility" of our reduced version of the CERAD battery consisting of only three subtests. Going back to the male patient's scores mentioned in

6 See Section §15 of Chapter 2, "Using SD for comparing the incomparable: Percents vs. *Z*-scores".

Section §35, we may ask how well or how precisely we are able to estimate his MMSE score y, given that we know his subtest scores $z_1 = 1$, $z_2 = -0.5$, and $z_3 = 0.5$. For a hypothetical example, a practical use may consist of the fact that while the MMSE can be carried out only by qualified MD psychiatrists, the three CERAD subtests could be administered by a laboratory assistant as a timesaving and auxiliary method for the following approximate estimation as the "*utility*":

Reduced version of CERAD test	Result of diagnosis
Composite $z \longrightarrow$ MMSE: y	
$z - r_{zy} \longrightarrow y$	
How well can we estimate?	

Recalling an analogy from elementary mathematics again, we can recognize that the test is treated here as an independent variable x from which we wish to estimate the values of the dependent variable y, which is now called a "criterion variable," or more succinctly, just the "criterion":

Independent variable – estimates → *Dependent* variable,

or, the same in test theory terminology

Test x – estimates → *Criterion* y.

From the basic properties of the correlation coefficient already mentioned above, we know that if the computed absolute magnitude of r_{zy} were very close to 1, we could closely estimate y (MMSE) from the known value of the composite score z. For practical use, however, this would mean that the error of estimating y has to be acceptably small. Then we can say that the composite test z has good validity with respect to y, and *the absolute value of the correlation coefficient* r_{zy}, $|r_{zy}|$, is "*the*" *coefficient of validity* by traditional test theory. Usually, when presenting the correlation coefficient r as a validity coefficient, its absolute value is automatically assumed, though very often this is not explicitly indicated.

3.2.1.2 Criterion-Related Validity

It is important to be aware that validity is always validity of something "*with respect to*" something, namely the *validity of a test with respect to an external criterion*. Or, the *validity of a test with respect to the decisions and actions that are being made based on the score*. About the criterion, it may be of a very different kind than the test instrument and its scores being compared. In some situations, the criterion is more or less *directly observable* or measurable by an empirical procedure, say by a clinical diagnostic; therefore, it can be called an *empirical criterion*. The related validity is then called *empirical validity*. Simply put, the "test" works here as an information source for estimation, while the "criterion" is what is to be estimated. In addition, since in our example, empirical validity is related to a criterion that is "outside the battery," we may specify it further as *external validity*. If the criterion is inside the battery, e.g. a subtest is to be estimated by the rest of the other subtests, then we can speak about *internal validity*. Thus, in such a case, a test may also be considered in the position of a criterion.

This usage of the terms external validity and internal validity is applicable when dealing with the validation of tests and measurement instruments. In experimental research studies, it has different meanings. Internal validity refers to whether or not the study design and execution were done properly so that the causal effect on an independent variable on a dependent variable can be shown and not affected by something other than the independent variable. External validity in a research study is about how the results apply and generalize in the real world. In Section §31 coming up, Petr discusses cause and effect in psychometrics and builds on that later to enable us to see the difference

between valid scientific variables that scale the levels of something real, and those that don't. It also enables us to appreciate the difference between "causal" and so-called "constructs."

If the estimation of an empirical external criterion is aimed at some practical purpose, as with the motivation to substitute one clinical method for another, such validity is sometimes called "*practical utility*," *See also Figure 3.1, the Evidential Basis row and Test Use column*. Whether this is done over a time interval for the purpose of prediction and/or more or less simultaneously, we can distinguish between *predictive* and *concurrent* validity, respectively. The scheme of predictive validity uses some additional specific terminology. The criterion is sometimes also called the *predictand*, which represents the dependent variable, e.g. academic performance, while the *predictors* are the independent variables, say admission tests.

The above-mentioned practical validity, or utility of the three-test battery represents a case of *composite validity* since we were estimating the criterion y (MMSE) using the agglomerated scores of the three CERAD subtests. For example, we might wish to sum up the z-scores for each patient. On the other hand, the simplest case of validity consists of using just one subtest, say the *C-Praxis* test, z_1. Then we speak about *simple validity*, denoted by or. The typical notation is especially when we work with the *raw* scores of a test x. The typical notation is by r_{xy}, especially when we work with the *raw* scores of test x.

3.2.2 (§27) Is the Error of Estimating the Criterion Sufficiently Small for Some Test Uses?

The value of a validity coefficient is crucial in assessing the "practical utility" of a given test. In psychometrics textbooks one can find a variety of recommendations on the problem of what value of the correlation coefficient is high or low. For instance, some authors say that validity greater than 0.80 is "very high" (and rarely found) that validity about 0.60 is "very good," that around 0.40 is "good," and that correlations of 0.30 are "low" but may still sometimes be useful, while those under 0.30 may or should be neglected. Therefore, rather than speculate about where the borderline is between "good correlation" and "very good correlation," it is better to use this classification only for a rough orientation, while "practical utility" should be expressed by the *standard error of estimation* of y, which can be assessed with respect to the precision for practical use of the corresponding regression equation as introduced in Section §28. Of course, in a given test use setting, there may be a number of situations that require accommodation to ensure that the tests are administered and scored in a secure, accurate, and standardized manner. Printed vs. computerized vs. tests requiring media displays introduce practical difficulties that may be more or less suitable in a given test use scenario, affecting their utility, and possibly introducing construct-irrelevant threats to validity.

The standard error $s_{y|z}$ of regression estimation of the criterion y from the test composite score z is given by:

$$s_{y|z} = s_y \sqrt{1 - r_{zy}^2}.$$

Intuitively, it might be considered a kind of "average error" of estimating criterion y based on knowledge of test score x. Recall from Section §28 that $s_{y|z}$ is a standard deviation (SD), namely s_e of the individual errors e of the estimates in each of the N individuals:

$$s_{y|z} = s_e.$$

Therefore, the rule of "6 sigmas" can also be applied here. Posing $\pm s_{y|z}$ around the estimated value of y computed by the regression equation, we may assume to get the 68th percent interval of precision. Usually, it is accepted that the interval $\pm 2s_{y|z}$ stands for the *maximum error* of estimate, corresponding to the 95th percent interval of precision.

3.2.2.1 Composite Validity

In our actual case, in the Utah study data, we have computed the unweighted (or equally weighted) average composite CERAD score z_c of the three subtests in each of the $N = 189$ patients. Then we computed the correlation coefficient between z_c and the MMSE points y, and we expressed the unweighted *composite validity*: $r_{yz_c} = 0.65$. We already presented the MMSE SD above to be $s_y = 5$. Thus, the standard error of estimating MMSE points, y, from the knowledge of the simple average z-score composite z_c becomes

$$s_{y|z} = 5\sqrt{1 - 0.65^2} \approx 5 \times 0.75 \approx 3.8 \text{ [points in MMSE]}.$$

Broadly speaking, this could be interpreted such that if you know the CERAD three subtest composite score z_c for a subject, you can estimate his/her MMSE points with $\pm 3.8 \approx \pm 4$ as an approximate "average" error of ± 4 [MMSE points]. Alternatively, in z-scores, you will estimate the position of a patient within the "average" error of ± 0.75 [z-score], or $\pm 3/4$ SD. It is clear that such imprecision in estimating the MMSE results is unacceptable, especially when comparing it to the maximum possible range of MMSE from 0 to 30 points. Thus, we conclude that the "practical utility" of the hopefully assumed substitution of MMSE by the shortened three-test CERAD would be poor and cannot be recommended.

3.2.2.2 Variability and Validity

In addition, the reader should learn from the example and the formula above that the precision and error of regression estimation do not depend *only* on the validity coefficient itself. They also depend on the variability of the estimated criterion variable in the sample, i.e. on its SD, the s_y in the formula. Imagine two theoretic situations with the same validity coefficient (0.65) but substitute the following numeric values into the formula above for two different cases, namely:

a) in a more homogeneous sample, say with half of the actual SD, then $s_y = 2.5$ would produce half of the actual standard error of estimation, $s_{y|z} \approx 2$, and
b) in a more heterogeneous sample, say with a 50% larger SD, then $s_y = 7.5$ would produce a 50% larger error of estimation, $s_{y|z} \approx 6$.

Warning. One of the terrible mistakes, and a confusing procedure that should *never* be applied, would be to use statistical hypothesis testing to find out whether the correlation coefficient of *validity* is "statistically *significant*." A "statistically significant correlation" actually means that the observed sample estimate of the correlation is "statistically different from exactly zero in the population," which heavily depends on the sample size N. Thus, for the typical significance level, $p = 0.05$, you will find that for a sample of $N = 100$, an $r_{zy} = 0.20$ is "significant." However, in a sample of $N = 1000$, even $r_{zy} = 0.06$ would be "statistically significant," meaning that it is not exactly zero in the population, but such a nonzero but very low validity has no practical or clinical utility. It is of no use for a prediction, and probably of little scientific importance in the behavioral sciences. The general problem of statistical vs. nonstatistical significance will be commented on in Section §29, under *Remarks on interpreting the magnitude of multiple R*.

3.2.3 (§28) Differential Weighting of Subtests: Composite Validity as Multiple Correlation

Now the question is: Can we improve the validity of a battery by cleverly choosing different weights for the different subtests? This involves a rational substitution of the unweighted composite by a weighted composite. If you still hesitate about the necessity of it, consider an argument from the opposite point of view: Can a researcher be sure that each of the subtests is of equal importance?

Is it reasonable to assume that the subtests are of equal importance even in the case that the battery is used to estimate different criteria? In addition, the possible unequal importance of subtests is well known from the very beginning of the history of psychometrics, and several rules of thumb have been suggested. Besides intuitive and expert-like weights, there is an approach analogous to the somewhat obsolete method of so-called item analysis. It recommends: "Give the subtests (items) weights that are proportional to their separate simple validity coefficients." If we were tempted to apply that out-of-date method, we would need to compute the individual validity coefficients between each of the CERAD subtests and the MMSE:

Simple validities of separate CERAD subtests to the criterion *y*, the MMSE.

Subtest	Subtest name	Validity
x_1	C-Praxis—"executive"	$r_{y1} = 0.59$
x_1	V-Fluency—"verbal"	$r_{y2} = 0.48$
x_1	L-Memory—"memory"	$r_{y3} = 0.52$

The most valid individual subtest, x_1 – C-Praxis, has a separate simple validity of 0.59, while in section §31, we found that the unweighted composite of all three subtests had a validity of 0.65. Thus, we have not gained much by adding the other two tests to x_1 and averaging their common score to make the unweighted composite z_c above. The so-called *incremental validity* of the three-test battery, produced by adjoining the two additional subtests to C-Praxis, shows an increment of just 0.06.

3.2.3.1 Weighted Composite Validity

Let's try to use the obsolete "item analysis" approach and multiply the three *z*-scores by the separate simple validity coefficients r_{y1}, r_{y2}, and r_{y3} in the role of the weights w_1, w_2, and w_3 for each subject:

$$z_{cw} = (1/3)(w_1 z_1 + w_2 z_2 + w_3 z_3),$$
$$= (1/3)(0.59 z_1 + 0.48 z_2 + 0.52 z_3).$$

For instance, let's use the illustration of a patient whose *z*-scores on the three tests were {1; – 0.5; 0.5}. Hence, the unweighted average composite score was $z_c \approx 0.33$, while his composite mean weighted by subjective expertise was $z_{cw} \approx -0.08$. Now, we get his composite z_{cw-r} weighted by the simple validities of the separate subtests, i.e. by the separate correlations r_j:

$$z_{cw-r} = (1/3)[0.59 \times 1 + 0.48 \times (-0.5) + 0.52 \times 0.5] = 0.20$$

We can see that this weighted method moves the patient's position even higher above the mean than the original unweighted method. Then, we first computed such a composite score for each of the patients in the rest of the sample in the same way. After that, we computed the correlation between this kind of weighted test composite z_{cw-r} and the MMSE criterion y to find $r_{yz_{cw-r}} = 0.68$. Again, the increment in validity, 0.03, when compared to the unweighted composite validity 0.65, is negligible. This minimal improvement is understandable as the three separate validity coefficients do not differ from each other very much. So, the weighted composite is not much different from the unweighted composite.

3.2.3.2 Maximal Validity Via Multiple Correlation

This attempt raises a question as to whether there is a general way of finding *optimal weights for maximum composite validity*. Yes, a method is available for finding the best possible differential weights to combine the *z*-scores of the subtests to get the mathematical maximum validity (that

is, correlation coefficient). This statistical method is known as *multiple regression*, and the maximal possible correlation related to it is the well-known *multiple correlation coefficient R*. However, it can also be formulated as a case of a simple "standard *r*," namely $r_{yz_{cw}}$ with an optimally weighted composite z_{cw}. Its usual notation by capital *R* may be further specified to indicate the variables involved, for example, $R_{y|123}$. Of course, the regression weights producing the highest possible correlation also yield the smallest possible *standard error of multiple regression estimate* $s_{y|123}$. There are different ways to compute the multiple correlation coefficient when the number of tests $n > 2$, in general, $R_{y|12\ldots n}$. Most of the methods were in use in the BC (Before Computers)-era, but certain formulas are also helpful for a substantive understanding of applications, and their relations to other multivariate methods, especially to the LCF model. For $n = 2$, the formula is so simple that if you find the few necessary input data elements needed for the computation, say presented in a research article, you may use a spreadsheet or smartphone calculator:

$$R^2_{y|12} = \left(r^2_{y1} + r^2_{y1} - 2r_{y1}r_{y2}r_{12}\right)/\left(1 - r^2_{12}\right).$$

In our example above, if we would like to compute the multiple correlation of $y -$ MMSE on just the first two tests, $x_1 -$ C-Praxis, $x_2 -$ V-Fluency, we already know $r_{y1} = 0.59$, $r_{y2} = 0.48$, and with the additional information on the internal between-test correlation $r_{12} = 0.38$, we get

$$\begin{aligned} R^2_{y|12} &= \left(0.59^2 + 0.48^2 - 2 \times 0.59 \times 0.48 \times 0.38\right)/\left(1 - 0.38^2\right) \\ &\cong (0.35 + 0.23 - 0.22)/0.86 \cong 0.36/0.86 \cong 0.42 \end{aligned}$$

Finally, $\sqrt{0.42} \cong 0.65$ (this rounded result completely agrees with exact computation). Thus, the maximum possible composite validity of x_1, x_2 to y is expressed by the multiple correlation $R_{y|12}$, enabling one to estimate $y -$ MMSE points, with the standard error of estimation $s_{y|12}$ as follows:

$$R_{y|12} \cong 0.65, \text{and } s_{y|12} \approx 4 \text{ [points on the MMSE scale]}.$$

It means that the optimally weighted composite of the $n = 2$ tests x_1, x_2 can offer the same validity as the unweighted composite of the simple sum of $n = 3$ tests x_1, x_2, and x_3 in Section §31, with $r_{yz_c} = 0.65$. A practical use of such a finding can be that the third test, $x_3 -$ Logic Memory, which is time-consuming, might be abandoned and considerable time saved without a substantial loss of information about MMSE. In reality, a study for standardizing an abbreviated test battery would need more detailed verification.

3.2.3.3 Multiple Correlation to Shorten a Test

The use of multiple correlation for *optimal test reduction* is important and useful. For example, think about substituting a field test for the diagnostic method in the role of a "gold standard." Nevertheless, besides the practical aspects just indicated above, there are some further connections to other methods, especially, for our purpose, to the LCF model (see later in Section §34 of this chapter). Note that from a certain point of view, the numerical result of the formula for computing the squared multiple correlation $R^2_{y|12}$ depends heavily on the sum $(r^2_{y1} + r^2_{y2})$, i.e. on the squared separate simple correlations of the external criterion y with each of the individual tests x_1 and x_2. Moreover, if these independent variables were uncorrelated so that the "information" brought in by the two tests were not "overlapping" as they are in the example, then the formula simplifies to just that sum. Hence, if $r_{12} = 0$, then $R^2_{y|12} = r^2_{y1} + r^2_{y2}$ since $R^2_{y|12} = r^2_{y1} + r^2_{y2} - 0$. The same holds for uncorrelated independent variables in the general case with $n > 2$.

However, one has to be very cautious in using the multiple regression method. Just blindly running the computational procedure of multiple regression analysis as it is offered by the usual

commercial statistical software can be very problematic. Especially, the automatic "stepwise" procedures may lead to nonsense. The reader is advised to learn much more about it from basic textbooks. Although the method is formally or mathematically optimal, it depends on the given sample, and on the given data at one's disposal. In addition, if a subtest is removed from the battery, or some new subtest is added to it, all the weights will change, sometimes very substantially. The regression weights express *mutual relative* proportionality of the tests/predictors specifically in a certain given "*context*" of the other independent *x*-variables. The topic of multiple regression can become very sophisticated and complex, and the *problem of* "*stability*" under different influences, not only "stability" in the sense of time or repeated measurements, is just one of the problem areas.

By common sense, it is clear that the set of differential weights of given subtests for estimating one criterion, say MMSE for diagnosing Alzheimer's disease, will change into another and a quite different set of weights when estimating another criterion, say to diagnose Parkinson's disease. Of course, this is given that the assumption that the use of the same battery with two different sets of weights was reasonable for distinguishing between two different criteria as it was presented in Section §26 on absolute validity vs. differential validity.

3.2.4 (§29) A Criterion Score is Best Estimated by a Particular Composite, Multiple Regression

Using any standard statistical software for computing standardized regression weights and multiple correlation, we obtain the following:

	Weights for "optimal" composite *z*-scores: *regression coefficients* (z-standardized)	**Maximum composite validity: multiple correlation, and standard error of regression estimate**
z_1 ... C-Praxis—"executive"	0.44	
z_2 ... V-Fluency—"verbal"	0.16	$R_{y\|123} = 0.71$
z_3 ... L-Memory—"memory"	0.30	$s_{y\|123} = 3.5$ [points]

We find that even the best weighting procedure is not able to produce sufficiently high validity, producing an increment of validity from the unweighted 0.65 to the optimally weighted 0.71. The standard error for estimating the MMSE criterion with an unweighted composite was 3.8 in MMSE points and 0.75 in *z*-scores. The smallest possible *standard error of multiple regression* is 3.5 in MMSE points, and 0.70 in *z*-scores, but is still not acceptable for "practical utility," say for practical estimation and substitution of the three tests for MMSE diagnostics.

3.2.4.1 Comparing Several Methods for Constructing Composite Scores

In general, we may consider constructing a composite score z_{cw} in many different ways. Up to now, we have seen three special subspecies of the general class of composite scores z_{cw}:

z_c: the special case with equal weights, each being just 1,

z_{cw-r}: the case with weights proportional or equal to the simple validities (correlation coefficients) of the separate subtests, the obsolete "item analysis weights," and

$\hat{z}_y$: the *regression-weighted score* constructed using multiple correlation *R*, the maximum possible validity based on this optimally weighted composite score.

Let us illustrate it for the patient with the formerly mentioned subtest scores $z_1 = 1$, $z_2 = -0.5$, and $z_3 = 0.5$. The "regression composite score" $\hat{z}_y$ is now constructed by the standardized *regression equation*: $\hat{z}_y = \sum w_j z_j$, which, in our example, becomes: $\hat{z}_y = w_1 z_1 + w_2 z_2 + w_3$[7]

$$\hat{z}_y = 0.44 \times 1 + 0.16 \times (-0.5) + 0.30 \times 0.5 = 0.51$$

by substituting the above-presented numerical values of the regression weights.

Although the regression equation for computing $\hat{z}_y$ is nothing more than just a variant of any weighted composite score z_{cw}, it is specific in its statistical conception and practical interpretation. The regression-weighted score $\hat{z}_y$ is the *best* of all possible weighted composites z_{cw} in the sense that it: (i) offers the best weights for estimating the criterion score y for every subject on average, (ii) $\hat{z}_y$ has the highest possible criterion validity and multiple R, and (iii) thus, using the $\hat{z}_y$ gives us the minimum standard error $s_{y|z}$ for estimating the criterion y.

3.2.4.2 Multiple Regression with Raw Scores

Usually, we are not interested in just the estimate of an individual's standardized value of the criterion; we would like to have not only the value of her or his criterion z-score but also the corresponding *raw score* estimate. In the case of simple validity, the bivariate regression in the form of raw scores can be directly expressed in the same way as in Section §28 and Figure 3.1, by the raw-score regression coefficients c and b. Then the estimates are produced in real-world units, say units of measured physical quantities like inches or centimeters, pounds, or kg, etc. In the case of composite validity with the weights established by multiple regression, the equation is extended to the form with *raw-score regression coefficients*:

$$\hat{y} = c + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n.$$

The raw-score weighted sum is a mysterious mixture of the individual observed variables in their original units of measurement. This mixture usually has no realistic interpretation and serves just as a mathematical mechanism for producing an optimal numerical value of the estimated variable y. Usual statistical software offers the values of raw-score b-coefficients with the intercept coefficient c, which allows constructing the whole equation for directly estimating the raw scores y.

3.2.4.3 Standardized Multiple Regression

However, we often prefer getting estimates $\hat{z}_y$ of y in standardized form. Then it is useful to know that the optimal regression weights w_j for z-standardized variables are usually denoted by β_j, $w_j = \beta_j$. Hence, the following regression equation gets the form with $c = 0$, and with regression coefficients called *standardized BETA-coefficients*:

$$\hat{z}_y = \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_n z_n.$$

These coefficients are typically printed under this name by most commercial statistical software. (Recall that we already met the simplest beta-coefficient in the section on bivariate regression, Section §27). However, we need not be interested in using the regression weights only for

7 Note that for estimating z_y, the regression procedure computes the weights w_j already multiplied by the averaging constant $1/n$, here $1/3$. Therefore, the equation becomes just a sum, not an average.

computing the estimates of the values of or for individual subjects. Sometimes we are focused directly on the weights themselves. It means that our primary interest would be in using them for computing a composite score of a test battery, as was introduced above.

3.2.4.4 Regression Calibration

Imagine a rather oversimplified situation where a battery of tests is being used for several years for admission of applicants. Assume that the battery has been validated repeatedly to have an acceptably high multiple R using its regression equation for predicting the final success of the applicants on a criterion y. Broadly speaking, the battery has been "calibrated" for a usual—perhaps every upcoming year—cohort of applicants. Then, it would not be a rational way of computing the total score of an applicant to use just the simple sum, or a by-experts weighted sum in the next, forthcoming cohort. Given that the equation shows reasonable stability then, by common sense, the best weighting procedure would use the beta-weights of the prediction equation to construct the overall composite score.

However, having constructed the standardized optimal composites, we may "de-standardize" them by using the mean and SD of the raw scores of y. In our example, the raw-score units of y are points on the MMSE scale. For instance, from the knowledge of a patient's z-score value, we can compute the patient's raw score simply by reversing the formula for standardization, into a *formula for "de-standardization"*: $\hat{y} = \hat{z}_y \times s_y + \overline{y}$. Using the above-presented mean and SD, $\overline{y} = 20$ and $s_y = 5$, we get $\hat{y} = 0.51 \times 5 + 20 = 22.601 \approx 22.5$ [MMSE points]. For an illustrative example of simplicity, let us use the roughly approximate value 22.5 rather than 22.601, since, in practical use, there is no rationale for considering MMSE to multiple decimals, yet perhaps a half-point difference still could offer some clinical discriminative information.

3.2.4.5 Error in Regression Prediction

Now, let us use the estimate of standard error of regression $s_{y|z}$, which we calculated earlier $s_{y|123} = 3.5$ [MMSE points]. From Section §12 of Chapter 1, we recall the application of the "rule of ±SD" to this standard error to see that for the interval $\pm s_{y|z}$, we have a 68% probability that our regression estimate will not jump over its bounds. Hence, we may expect that having estimated a patient's score as 22.5, the real MMSE score will be roughly within the interval $(\hat{y} \pm s_{y|z})$, i.e. (22.5 ± 3.5), which means somewhere between 19 and 26. Thus, we can conclude that such an error indicates that the external empirical validity of the three CERAD subtests for estimating MMSE as the target criterion is not high enough, and probably, of no "practical utility."

In connection with the interpretation of validity in terms of "practical utility," let's recall the possible intuitive view of the SD s and the standard error $s_{y|z}$ as numbers that resemble the "average" by computation: The average of squares later seemed to be "recomputed back" by taking the square root. Thus, on one hand, considered in this intuitive way, the standard error may seem to be a kind of "average error."

3.2.4.6 Remarks on Interpreting the Magnitude of Multiple *R*

1) *On statistical significance of R*: Usual statistical software automatically offers information about the statistical significance of R. Since here "statistically significant" means "not zero in the whole population," it would be nonsense to interpret a "significant R," i.e. "nonzero validity" as "a good validity." First, in the case of a very large sample size N, even a very small R might be "significantly" different from zero. Second, in a usual validity study, the sample is not a

randomized representative sample of any population. Third, the main informative index about the practical use of validity is the standard error of the estimate, $s_{y|z}$.

2) *On the number of independent variables, n, and its ratio to the sample size N*: Even in the case when we are interested in just a given sample of people, there is a problem known as *capitalization on chance*. It becomes more influential if the number of predictor variables is large. A part of the problem deals with a "maximization effort" of the mathematical technique of regression: Unreliability, all measurement errors, and slight contextual influences at the moment of data collection are subsumed into the equation—a "*vacuum cleaner effect*"—to construct a maximal possible value of *R*. Furthermore, not many authors warn researchers about the problem of "the number of subjects vs. the number of variables." For example, it is well discussed from a realistic position for research practice by Schutz and Gessaroli (1991). However, some other very rigorous authors recommend that the $N : n$ ratio should be as high as 40 : 1 (Cohen & Cohen, 1975). This ratio problem can be illustrated by two examples.
   i) With just one independent variable x ($n = 1$), one dependent variable y, and $N = 2$ subjects, the raw score data matrix of the independent variable values, $N \times n$, becomes 2×1, just a column. Hence, there are just two subjects' points in the (x, y) system of axes, and the problem of fitting a "best" regression line degenerates to just one possible line given by the two points with correlation. From a general point of view, this was a special case where, and by analogy, in multiple regression, whenever it holds, the multiple R would also be 1.00.
   ii) A further related problem is that the maximizing tendency of multiple regression will almost always construct an $R > 0$ just by chance. The chance increases as $(N - 1) \longrightarrow n$. It is possible to identify what value of R we may expect by chance for a given ratio of *Subjects : Predictors*. The value of R^2 is n/N (see e.g. Montgomery & Morrison, 1973). Then the computed R^2 should be compared with this value and taken with this caution—as exemplified and criticized by Armstrong (1975) in his "*1973 Tom Swift Award for Data Abuse*." If the ratio $N : n$ is low, then multiple regression yields a deceptively high multiple R, falsely indicating good validity. He gives an example of a nonsensical published paper predicting medical exam scores with $n = 14$ and $N = 18$. In our example with $n = 3$, and $N = 164$, the ratio is $3/164 \cong 0.02$, which seems negligible when compared to the computed $R^2 = 0.506$. Then this value adjusted with respect to 0.02 yields the adjusted $R^2 = 0.497$, not substantially lower than the one computed.

3.3 Explanation Requires More Than Correlation

In this Subchapter 3.3, Professor Blahuš is continuing his treatment of the role of multiple regression in scientific explanation. He immediately discounts regression, and hence correlation, which is a standardized regression coefficient, as providing any fundamental type of explanation. He calls it "a statistical metaphor." Recall that the *validity argument* theorists place at the center of validation the *interpretation* of a score from some measurement instrument in the context of *what action is to be taken*. Test takers have the right to a sound explanation of what their score means, and how, along with cut-off scores of some kind, it merits the action of failure or success on some goal of importance to them. As the balance between student and institutional control keeps shifting toward the student, more and better explanations are needed, and thus greater internal validity evidence that the instrument is truly valid for the intended action. Scientific explanation is truly increasingly

essential. What is better than causal evidence? In medicine, clinical trials give a drug and a placebo to different groups from the same population of sufferers from some condition the drug may improve. Comparison with one or more control groups gives credence to the hypothesis that the treatment had a causal effect manifesting as improvements in the experimental groups. A drug cannot be approved using correlations found through epidemiological studies alone.

Starting with Section §30, we will note a shift from the easier to understand external, or criterion-related aspects of validity, to internal aspects. Of Messick's first three internal aspects of construct validity ((i) Content, (ii) Substantive processes, and (iii) Structure), substantive processes are the aspect of construct validity most closely related to causal explanation. Substantive means filled with relevant substance, and with processes of thought or body movements deeply theory-connected. Causation is the most respected form of construct validation evidence. This kind of evidence can establish both scientific validity and real-world meaningfulness, providing a grounding for clear explanation.

3.3.1 (§30) *Scientific Explanation and Prediction*: Multiple Correlation and Regression as Its Statistical Metaphor (i)

Regression is often one of the two main types of statistical models that serve as a parallel or metaphor for scientific explanation. Regression, considered *explanatory metaphor* (i), is metaphorically 'explaining' just a certain, separate, *single value* of *a single variable* by producing an estimate or prediction of *criterion variable* based on the regression equation. For instance, it uses a certain test score for a subject to estimate predicted success or performance, a clinical diagnosis, suitability for a type of job, etc. On the other hand, another statistical parallel of explanation, that is, *explanatory metaphor* (ii), is a metaphor for "explanation" of a *relationship* of *two or more entire variables*, not just an observed individual value of one variable. In a more complex data set, such as a typical spreadsheet file, the "entire" relationship may be understood as a relationship between two or more columns of values. We are contrasting metaphor (ii) with metaphor (i), which was focused on estimating only a single and separate value on one variable.

Since regression and correlation by themselves do not represent causal relationships, this statistical metaphor reduces *explanation to merely a descriptive estimation or prediction*. This statistical metaphor could become somewhat closer to genuine explanation were it to fulfill several minimum conditions mentioned in Sections §31 and §36, where we discuss the problem of correlation vs. causation. The reader may be familiar with statistical literature that uses the term "explanation" in a way which is not appropriate from the point of view of general methodology of science or philosophy of science and their perceived view of the principles of causality. For example, terms like 'explained' variance should more correctly be called "estimated variance" and so on. To keep minimal consistency with other texts, we will still use this kind of terminological jargon but express it using the special single quotation marks ' ', as in '*explained*.'

3.3.1.1 "Explained" Variance in Regression

The formal statistical power of the regression approach—the fitting of a regression model—is expressed through correlation coefficients. This includes the multiple correlation coefficient, $R_{y|12...n}$, which could also be symbolized as $R_{y|z}$ to underscore the special case where x_1, x_2, ..., x_n are optimally combined into a composite z. Another expression of the regression approach, formally equivalent to the correlation, is based on the idea of expressing the fit by the *proportion of* '*explained*' *variation* in the dependent variable y. The decrement in uncertainty regarding the location of the predicted scores y' could be expressed by the proportion of the reduced variance of

the error of the $\boldsymbol{e}$ estimate $s_e^2 = s_{y|z}^2$. This is assessed as a ratio with respect to the original variance (as 100%) $s_e^2/s_y^2 = s_{y|z}^2/s_y^2$ and is referred to as the "unexplained" proportion of variance. The "explained" complement is $1 - s_{y|z}^2/s_y^2$, the proportion of variance accounted for.

Now, recall the formula for the standard error of regression estimate, $s_{y|z} = s_y\sqrt{\left(1 - r_{zy}^2\right.}$ from Section §27. By squaring both sides, and expressing $r_{zy}^2 = 1 - s_{y|z}^2/s_y^2$, we see that the "explained" proportion of variance can be computed more practically as the square of the correlation coefficient. This square is called the *coefficient of determination*, and in our example of composite validity, the multiple correlation $R_{y|123} = 0.71$ gives the coefficient of determination as $R_{y|123}^2 = 0.511$. Thus, we may conclude that just 51% of the original variance of the dependent variable MMSE has been "explained" or "determined"—properly speaking is *estimable*—by the three CERAD subtests. The rest, 49%, is left to some other relationships, mostly of an unknown nature. This complementary variance ratio, $s_{y|z}^2/s_y^2$, is then known as the so-called coefficient of *nondetermination*, or *alienation*, $1 - R_{y|z}^2 = 0.49$, which expresses the 'unexplained' proportion as indicated above correlation gives the coefficient of determination 1.

3.3.1.2 Assessing Imprecision in Regression

By intuition, it seems that instead of using the coefficient of nondetermination, and instead of using proportions of variances, we may use the *ratio of the SDs*: that is the square root of nondetermination. Using this SDs-ratio has a practical interpretation, which corresponds better to the common sense of *imprecision*. Following this kind of common sense, and brought in by norms of technology, the *relative error* known as *imprecision* is expressed by the ratio of a maximum error to the range of measurement, with the ratio usually in percent. However, by the rule of 6 SD (see Section §16), the range equals approximately to $6s_y$, and the maximum error to $6s_e$, and since in a ratio the constant 6 vanishes, we may formulate the "*relative imprecision*" *of regression estimate* by, or [%], or [%]. For instance, the predictive utility of a test z is given by the standard error of estimation of score y in measurable units, in the example of Section §29, this is $s_{y|z} \approx 3.5$ [MMSE points].

Now, imagine that before carrying out the regression analysis, we know *nothing more* than that a certain subject is a randomly selected member from a sample whose mean and SD are known: $\overline{y}$, s_y. Then, under the assumption of the $\boldsymbol{N}$-law, our one-valued "best guess" of the unknown score y would be $\overline{y}$ for that subject, with an uncertainty that the correct score can be somewhere in the interval $(\overline{y} \pm s_y)$ with 0.68 probability. (This is just another look at the same problem as at the end of Section §28, now from the opposite perspective.) Such a guess for a patient from the Utah survey would be 20 ± 5 [MMSE points]. Nevertheless, if we have some additional information due to nonzero correlation with CERAD, and knowing his/her scores in the three CERAD subtests, we may compute his/her regression estimate $\hat{y}$, and compare the standard error of this regression estimate, $s_{y|z} = 3.5$, to the SD of the preliminary guess $s_y = 5$ [MMSE points]. That proportional *decrement in the error* may be interpreted as the "amount of information" brought in about the estimated exact value: $1 - s_e/s_y = 1 - 3.5/5 = 0.30$, certainly, it corresponds to $\sqrt{1 - r_{xy}^2} = \sqrt{1 - 0.51} = 0.70$, which means a 30% *smaller error*, i.e. a thirty percent better estimate. However, from the opposite point of view, it means that the "imprecision of the estimate" is still 70%, to use the analogy with measurement in technology.

Warning 1 on interpreting "explained percents"—distinguish between (i) "explained" variance of y vs. (ii) "explained" error/imprecision of estimating y.

For a better understanding and appropriate interpretation of research results, it is quite useful and important to be aware of the difference between the meanings of the two notions. If we accept the relative decrement of the (technological) imprecision, the SD ratio, s_e/s_y, as an analog of "explanation" to the traditional statistical approach using the variance ratio, s_e^2/s_y^2, then compare: $s_e/s_y = 0.70$ vs. $s_e^2/s_y^2 = 0.49$. In traditional statistical terminology, 49% of the variance of the dependent variable y "was left 'unexplained'." However, the relative *imprecision* of estimation of the y values is still 70%. Thus, contrary to usual statistical habit, it seems that in some practical situations, it is more informative to measure the percentage of "explanation" by the *relative "imprecision" of estimation* of a value of y in its original measurement units but not in their squares. Moreover, the tradition of speaking about *explained percents* in terms of the coefficient of determination ($r^2 \times 100$) without any further specification may falsely evoke the impression that it deals automatically with the error of estimation in the sense of technological measurement, while it actually expresses a ratio of variances. In addition, note that the ratio of variances is necessarily lower than the ratio of SDs—due to simple arithmetic. This may distort a subjective assessment and falsely overestimate the quality of a diagnostic method, its validity, or predictive utility for practical use.

Warning 2 on "explained' percents"—distinguish between: (i) "explained" variability of y vs. (ii) "explained" magnitude of y.

It is not a rare case, unfortunately, when clinicians, medical doctors, and even researchers misinterpret the "explained" percentage by saying things like these:

- The value of *IQ* is inherited up to (?) percent.
- Body height of a child can be "explained" up to ... percent by the average height of the parents.
- The long jump performance of the former Czechoslovak record holder Gombala can be "explained" and predicted up to 25% by the value of his running speed at the moment of his take off, etc.

The above-presented percents in this *Warning 2* were established based on regression in terms of the coefficient of determination $r^2 \times 100$ [%]. Therefore, the last proposition should have been reformulated more properly as follows: The *variability* of repeated performances y of Mr. Gombala over different competitions can be predicted up to 25% of their variance, $s_y^2 = 324\,[\mathrm{cm}^2]$, by knowledge of his running speed x. This athlete was measured many times over a period of time, and the mean of all the athlete's jumps over the period was $\bar{y} = 760$ [cm], and the SD $s_y = 18$ [cm]. Simultaneously with the jumps, his running speed x at the moment of take-off was also measured and his personal regression equation was established for estimating his performance y from the running speed x. It yielded the correlation coefficient $r_{xy} = 0.50$, i.e. $r_{xy}^2 = 0.25$—it means 25% *of the variance*, i.e. 25% of 324 [cm^2]. The standard error of regression estimate of his jump y was $s_{y|x} = 15.6$ [cm], which does not seem to be much better than the original SD of his performances, $s_y = 18$ [cm]. In terms of the error of estimate, the "imprecision" was reduced or "explained" up to $(s_y - s_{y|x})/s_y = (18 - 15.6)/18 \cong 0.13$, which means 13%—the fact that this is lower than 25% in terms of variances is related to the above-mentioned *Warning 1*.

However, the most important thing to be learned in the sense of this *Warning 2* is: The percents are related to the variability in a range of about $\pm 3s_y = \pm 54$ [cm] around the mean $\bar{y} = 760$ [cm]—but the percents must not be applied to the whole magnitude, say $y = 760$ [cm]. It would be *nonsense* to say something like "25% of the 760 [cm] jump is due to the speed of running." Thus, it would be nonsense to conclude that 190 [cm], i.e. 25%, was due to the speed of running—and, perhaps—therefore the rest of the distance, 570 [cm] could have been made in terms of a standing long jump. Professor Blahuš says, "I dare present this example here only because I have witnessed exactly this

misinterpretation (cf. Blahuš, 2000)." Similar percent misinterpretations are not rare in medicine, genetics, etc. Hence, the moral of *Warning 2* is that $r^2 \times 100$ [%]—primarily expresses percents of variance, s_y^2, i.e. of a number whose physical meaning is in *squares of measurement units* of the original observed values y, and secondarily, this number (s_y^2) is a characteristic of variability, hence its percents may be related only to the *variability* of y but *not to the magnitude* of y.

3.3.1.3 Fisher's Information and Efficiency Measures

Statisticians mostly prefer to use the proportion of variances, and percents of variance, and speak about "percents of 'explained' variance." Maybe, this habit is due to the elaborated inferential theory with the well-known Fisher-Snedecor distribution function for the F-statistic. It allows assessment using the F-test, a test of the ratio of two variances created by R. A. Fisher, the father of the analysis of variance (ANOVA) early in the 20th century. In connection with this, it seems interesting that historians of statistics do not sufficiently appreciate Fisher's concept of "information" which he formulated around 1921, several decades before the era of cybernetics and information theory. Fisher had considered variance of errors a measure of "uncertainty," therefore *Fisher's information*, I, is defined as the opposite of the "uncertainty," namely the *reciprocal of the variance of the errors* of estimation, $I_{y|z} = 1/s_e^2$ (N.B.: Again, it is not based on a standard error but on error variance, meaning it is based on squared units of measurement). In this sense, knowing the additional "message" about a given individual's CERAD score z, the corresponding amount of information becomes $I_{y|z} = 1/s_{y|z}^2 = 1/3.5^2 \cong 0.08$. However, if we knew just the complete variance of the whole Utah sample, we have rather little information for estimating the MMSE score y, namely $I_y = 1/s_y^2 = 1/5^2 = 0.04$. Note that I is expressed in real (squared) measurement units—here reciprocals of the squared MMSE points, [MMSE points]$^{-2}$. Therefore, to ensure comparability, a ratio of two Is is used, called relative efficiency—or just *efficiency*: $Ef = I_y/I_{y|z}$. Here $Ef = 0.04/0.082 \cong 0.49$, which is equal to the coefficient of nondetermination (alienation) mentioned above, $Ef = I_y/I_{y|z} = s_e^2/s_y^2 = 1 - R_{y|z}^2$ (the ratio of the two Is equal to the reciprocal of the ratio of the variances). The concepts of information and efficiency are being used in different contexts in statistics as well as in test theory and psychometrics. In the present context of regression estimation, we should note that from the last formula, we may also get the *coefficient of determination* $R_{y|z}^2 = 1 - s_e^2/s_y^2 = 1 - Ef$. This complementary concept to efficiency thus suggests an interpretation of R^2 as a kind of "Relative Information," $1 - Ef$, whose advantage is that it is a unitless number from 0 to 1.00 and may be compared over different variables with different measurement units (inches, seconds, etc.). The main power of multiple correlation and regression consists in the assessment of "diagnostic information"—i.e. *validity*—of the tests. However, its use for estimation and prediction may hardly serve to support a *genuine causal explanation*. The usually desired requirements posed on the causal explanation are presented now.

3.3.2 (§31) Correlation vs. Causation: Their Respective Roles in Scientific Explanation

In the general methodology and philosophy of science, several types of scientific explanation are usually recognized. Among these, it is the paradigm of *causal explanation*, which usually deserves the main attention and highest recognition as a device of the scientific approach to understanding real-world phenomena. One of the main features of its explanatory power is that actively producing a *change* in the value of the independent variable—the *cause*—produces a change in the value of the dependent variable—the *effect*. This feature is one of the properties, which explains the main

conceptual difference between two general methods of empirical research, viz. *observation* vs. *experimentation*.

In laboratory jargon, a session where tested subjects undergo a procedure of measuring their reaction time might be called "an experiment." However, from the point of view of methodology, such a procedure itself represents just a case of observation. The simple separated procedure of measurement itself is nothing more than quantified observation. Administering a questionnaire, applying a battery of tests, etc., are usual instances of empirical scientific methods of observation. In addition, we must not confuse this meaning of meticulous observation as a general methodological approach with casual "observation" by sight, as a sensory process.

3.3.2.1 Experimental Manipulation and Passive Observation

What makes the methodological difference of an *experiment* is neither complicated organizational conditions nor sophisticated technological devices in a laboratory. Its specification is conceptually best represented by an actively influenced system's behavior in terms of the paradigm of Input-Output or by the Stimulus-Reaction (S|R) analogy: Does an active change Δx in x produce a change Δy in y? A very plausible case occurs when we are able to control the change of x by our intentional manipulation. A rigorous view is sometimes represented by the maxim "No proof of causation without manipulation." Behavioral and social sciences, however, very often deal with descriptive data, like survey data, that are not influenced by intentional manipulation but only observed under spontaneous natural conditions. Such data are a prototype for which the well-known statement is valid: "Correlation does not mean causation." However, this does *not* necessarily apply to the correlation between the change or increment Δx in the manipulated variable x in a position of an experimental factor, and the produced change Δy in the effect variable y. Take, for instance, the correlation between the amount of a drug and the increment of errors in a behavioral task.

A typical counter-example is the correlation between smoking and lung cancer. If the correlation is observed in different subjects under their spontaneous conditions without manipulation, we do not have any rigorous methodological reason to assume causal influence. Methodological support may be brought in by an experiment with manipulated doses of different amounts of cigarette smoke applied to laboratory animals. On the contrary, no such similarities might be imagined for the correlation of body weight with body height although this correlation is quite high. Body height cannot be manipulated in adults at all. Some manipulation of body weight by overeating or bodybuilding exercise is possible but we will certainly reject the idea that these could produce an incremental change in body height. Moreover, we would reasonably hypothesize that—of the two—neither the first nor the second is a cause or an effect of the other, but both are joint effects of their common causal complex of genetic variables, nutrition variables during early ontogeny, etc.

Therefore, we have to distinguish *two types of correlation*. On one hand, we have a "passively observed," descriptive correlation as an unspecified association under *spontaneous co-variability* of the status quo states of x and y under "natural" conditions. On the other hand, there is a correlation between manipulated increments Δx and the increments Δy *produced by the manipulation*. Graphically, this is the difference between

$$x \longleftrightarrow \text{y vs. } \Delta\text{x} \Longrightarrow \Delta\text{y}.$$

An example suitable for illustrating this difference may be the consideration of the two different types of correlation between the amount of time x spent on learning and the performance y in a test of academic performance. On one hand, we may find a correlation r_{xy} by asking students how much time they spent on learning and then giving them the test. On the other hand, we can apply a pretest first, then we instruct and "force" the students to spend different prescribed amounts of time on

learning, and, finally, we correlate the time increments with the increments of pre-test vs. post-test scores, obtaining thus $r_{\Delta x \Longrightarrow \Delta y}$.

3.3.2.2 The Counterfactual Condition and a Thought Experiment

In some instances, as in the latter example, we do not have data obtained by controlled experimentation, but the manipulation may still be imagined at least in an abstract, ideal experiment in the researcher's mind. Of course, such a thought experiment is not sufficient for methodologically jumping over the barrier dividing an observational association from causation. However, it represents one of the minimal necessary conditions, which should be checked at the very beginning of considering an observational correlation to be a hopeful indication of a *potential causation*. As a check of this minimum assumption, we might apply the following advice: "Imagine the *counterfactual condition* of manipulation vs. nonmanipulation." It means: What would have happened to the dependent variable if the independent variable had been manipulated to take on a different value? Is it even possible, at least in principle? Is a manipulation even thinkable to be carried out through other variables? The answer "yes" is not sufficient, of course, but it represents a necessary minimum condition for any reasonable hypothesizing that the investigated relationship could potentially serve as a basis for causal explanation.

The ideal thought experiment, which would carefully check the real practical possibilities of manipulating each of the independent variables in a multiple regression equation, is the minimum condition for serious consideration of the descriptive regression as a virtually hopeful *statistical model (i)*—or just a metaphor—of possible scientific explanation. Thus, not only is a synchronic random coincidence different from causality but so also is a diachronic relation, formulated by the maxim "*Post hoc ergo propter hoc*," "After, therefore (causally) due to," as a prototype of a mistakenly interpreted predecessing and succession as a cause-effect relation.

3.3.3 (§32) The "Cleaned Off" Incremental Validity of a Subtest: Partial Correlation

In the preceding Sections §28 and §29 on weights of subtests, we saw that even the best weights assigned to subtests, including the additional subtests, need not produce remarkable incremental validity. Recall the case of three mentioned regression weights creating the best composite validity $R_{y|123} = 0.71$ of the three-test battery to the external criterion y (MMSE). It was not too much better than the former simple validity $r_{1y} = 0.59$ of only one subtest x_1 alone (executive). After having a look at the other coefficients of validity in Table 3.1, this finding can be understandable by common sense as follows. On one hand, the two added tests, x_2—verbal, and x_3—memory, have quite high simple validity to the criterion: $r_{2y} = 0.48$, and $r_{3y} = 0.52$. So, it seems that each of them *does* contain some relevant information about the *external* criterion y. On the other hand, the two tests have quite high *internal* validity to the former x_1—Executive. These intra-battery validities of the x_2, x_3 subtests to the former x_1, cf. $r_{12} = 0.38$, $r_{13} = 0.32$, may be interpreted in the way that each of the two new subtests contains a part of the same information about y as the original subtest x_1 already did. So, they are "overlapping" with x_1 and therefore do not bring much *new* information.

3.3.3.1 The Dilemma of Internal and External Validity

Here is the dilemma: If one message informs you about an event, the other message must be related to the first—otherwise it would not speak about the same event. On the other hand, the second message should not be related to the first one so much that it would become almost identical to it—otherwise it would not tell you anything new. Therefore, if the newly added subtests have

Table 3.1 Internal and external validity.

		Internal validity r_{jk}**, mutually among the subtests: coefficients in their correlation matrix** $\boldsymbol{R} = [r_{jk}]$			
		x_1	x_2	x_3	
	x_1—Executive	1	0.38	0.32	
	x_2—Verbal	0.38	1	0.51	$= \boldsymbol{R} = [r_{jk}]$
	x_3—Memory	0.32	0.51	1	
External validity:	y—MMSE:	0.59	0.43	0.52	r_{yk}
		Correlations of subtests x_j to the criterion y			

too high (internal) validity to the former subtests, they would not improve the former external validity of the battery too much—they will produce low *incremental validity* for extending the battery. This problem can be partly solved by a well-known compromising *rule*: "Choose and add new additional subtests that have relatively high validity to the external criterion while having relatively low (internal) validity to the formerly chosen subtests."

Thus, a question may arise: How can we express the separate or *partial contribution* of a new subtest? We need to establish the contribution that is "cleaned off" from the information the new subtest already shares with the other subtests. This is part of validity that could be regarded as a "clean validity" or "partial validity" after the overlapping information has been "partialed out." The answer to the question is offered by an additional multivariate method, the coefficient of *partial correlation*.

3.3.3.2 Partial and Spurious Correlation

Let us make a short side step to anthropometry. A physical anthropologist measuring the length of the forearm and the length of the foot will find a very high correlation between them. However, after having correlated these two with body height, he/she will find that the two strongly depend on it. Some researchers speak about *spurious* correlation due to the influence of a third variable, which works as a common "cause" of the two original ones. Then, what would be the "cleaned off" correlation between the forearm and foot if we were able to eliminate the "influence" of body height? This is what the partial correlation is able to tell us merely by using a few calculations: "What the correlation would have been if the third variable had been constant."

Now, let's have a look at a possible sequential approach to constructing a battery by a one-by-one addition of new tests to the first one. We already know that the subtest x_1—Executive has the highest separate simple validity to the criterion y − MMSE, namely $r_{1y} = 0.59$. We may decide to choose as the next added test the one with the next highest separate simple validity. This is x_3—Memory test with its $r_{3y} = 0.52$. But, having already selected x_1, what would be the "cleaned off" validity of the new x_3 after eliminating the portion of information that it shares with the previously used x_1? We find the answer by using standard statistical software or a quite simple formula and a calculator to compute the *partial correlation* between x_3 and y with x_1 "partialed out": $r_{3y|1} \cong 0.43$. In our example, we can compare the simple correlation $r_{3y} = 0.52$ to the partial correlation $r_{3y|1} = 0.43$, which is lower due to its mutual internal validity with x_1 ($r_{13} = 0.32$). The decrement to the cleaned validity, $0.52 - 0.43 = 0.09$, indicates the "overlap" between information about y brought by the

new subtest x_3 and the information that was already contained in x_1. The not negligible cleaned validity 0.43 with the small overlap 0.09 suggests that the subtest x_3 is not overly redundant.

3.3.3.3 Calculating Partial Correlation

The formula for calculating the partial correlation coefficient is not only simple but also conceptually useful for understanding another would-like-to-be "explanatory" role of correlation including in latent variable modeling. Therefore, the reader is invited to repeat the numerical calculation, even with the roughly rounded numbers from Table 3.1.

$$\begin{aligned} r_{3y|1} &= \left(r_{3y} - r_{31}r_{y1}\right)/S \\ &\cong (0.52 - 0.32 \times 0.59)/0.76 \cong 0.33/0.76 \cong 0.43 \end{aligned}$$

where the number $S = \sqrt{(1 - r_{31}^2)\left(1 - r_{y1}^2\right)} \cong 0.76$ in the denominator may be regarded as just a scaling factor since $r_{3y|1} = 0$ *if and only if* the *numerator* $(r_{3y} - r_{31}r_{y1}) = 0$. If this were the case, it would mean that the validity r_{3y} of the additional subtest x_3 is fully overlapping with the former subtest x_1 yielding no new information about y. Their difference may be used as an approximate "*numerator check*": $(r_{3y} - r_{31}r_{y1})$. This offers fast information based simply on bivariate correlations: If the product $r_{31}r_{y1}$ were roughly equal to r_{3y} then the incremental validity of the new additional test x_3 would be very low.

Here is also another useful look at the partial correlation. In Section §28, we defined regression in terms of $\overline{y}|x$, the conditional mean of y given that x is fixed. In a similar way, we are interested here in $r_{3y|1}$ in the position of a "conditional correlation" between x_3 and y given the condition that x_1 is fixed or "located."

3.3.4 (§33) Scientific *Explanation of a Relationship* Between Attributes—Between Whole *Variables*: Partial Uncorrelatedness as Its *Statistical Metaphor (ii)*

The statistical metaphor (i) in Section §30 dealt with using multiple correlation and regression for "explanation" by estimating only *one value* of just *one variable*. Now, the second statistical metaphor (ii) of explanation uses partial correlation to "explain" *relationships between several variables*, for example, a correlation between two psychological tests and with a third variable. But, how far such an 'explanation' with correlational dependence of the two on a third variable may be assumed actually *causal* must be taken only as a parallel or metaphor to the genuine explanation outlined in Section §34. Let's go back to the anthropometric example of the correlation r_{12} between the length of the forearm x_1, and the length of the foot x_2 introduced in Section §32. The correlation is illusory where each of the two variables is highly correlated with a third variable, v, namely the overall body height, which, simply using common sense, might be their common "cause." The very high correlation r_{12} practically vanishes after having the "influence" of body height "partialed out" or "located" and we find $r_{12|v} \approx 0$. This could be accepted as a good formal index for saying that the original direct relationship $x_1 \longleftrightarrow x_2$ has been "explained" by the joint dependence of the two on body height v.

Some typical spurious correlations due to a third variable have been found in past school surveys that were trying to find variables "determining" *IQ* in children. Due to the usual dispersion of age of the children in the samples, such studies might find many nonsense correlations between *IQ* and any of the other variables that depend on the general level of overall development and growth of the children. Age works here as an omitted variable indicating the child's overall development. As an exaggerated illustration, imagine a heterogeneous sample of children with a substantial range of

age, say from 8 to 15 years, and you would find a high correlation between *IQ* and body height, forearm, foot, as well as any other anthropometric measure, say the length of the left thumb.

3.3.4.1 Partialing Via Sampling Stratification

The idea of "partialing out" the correlation between an x_1 of IQ, and an x_2 of body height—due to their common covariation with age—v may be illustrated if you imagine a selection of a subsample of children of exactly the same age. Assume that we had found $r_{12} = 0.60$ in the sample of $N = 10{,}000$ children with age v ranging from 8 to 15 years. Then, imagine that we selected the children with the exact age of 15 years and 365 days, say there could be 100 such children, and compute the correlation between x_1 and x_2 just in these hundred children homogenous with respect to age v. This correlation would be much lower or almost zero, of course. Next, we select the subsample of children aged 15 years and 364 days and compute the correlation in it, etc., lower and lower, day by day, to the age of 8 years. Most likely, we would find that the correlation in each of those subsamples is almost zero. Suppose that the average correlation over all these subsamples actually was zero, which we designate by $r_{12|v} = 0$. In such a case, we might say that age perfectly "explained" why IQ and body height were correlated. So, when we used a fixed age, i.e. when we kept variable v at some *local value* v^* then IQ and body height—the variables x_1, and x_2—became uncorrelated at this value, also said to be "locally uncorrelated," or conditionally or partially uncorrelated.

3.3.4.2 Partialing Via Statistical Control

To strengthen our critical view of causal misinterpretation of the post hoc descriptive correlations still further, let us choose a developmental variable x_2, which is accessible to experimental manipulation: x_2—*Grip Strength Test* of the right hand, measured by a dynamometer in [kg]. Now let us consider the correlation between $IQ - x_1$, and *Grip Strength*—x_2 in an artificial example with fictive data as shown in Figure 3.3. Just by comparing the two left columns of data, we can easily see the strong relationship between *IQ* and *Grip Strength*, even without knowing their computed $r_{12} = 0.75$. Does it mean that by implementing an effective strength-training program, specifically for the right hand, we could increase the *IQ* of the ten students? By looking at the third column with age data, we find that the spurious correlation is due to the general developmental effect and that the third variable, *Age*—v, is "explaining" the spurious correlation almost completely to zero. It is indicated by the *residual* partial correlation 0.03 as well as by the approximate rough *numerator check* 0.01 (cf. preceding Section §32). This may also be graphically illustrated in Figure 3.3.

Of course, if we intend to explain a relationship between two variables, we would wish to "explain" the correlation perfectly. So, we would like to find such an additional "explanatory" variable or variables so that their partialing out would bring the partial correlation of the original variables to zero. This means that after "localization" of the values of the additional "explanatory" variable, the correlation of the original variables has been completely "explained."

3.3.4.3 Local Uncorrelatedness and Falsifiability

Thus, in the statistical metaphor *(ii)* to scientific explanation, this requirement represents the scientific need for *completeness of explanation*. Its parallel expressed in statistical terms may be called the *principle of partial uncorrelatedness*, or *local uncorrelatedness*, due to the above-mentioned elimination of covariation with the "explanatory" variable by making its values fixed, or "localized." In research practice, the requirement of completeness cannot be absolutely fulfilled. Therefore, it reduces to the more realistic demand that the "unexplained" *residual correlations* be negligibly

Fictive data for 10 students

Student	*IQ*	*Grip strength*	*Age [years]*
	x_1	x_2	v
1	93	20	13.0
2	95	28	13.0
3	98	22	13.4
4	100	21	13.8
5	101	41	13.9
6	104	50	14.0
7	109	38	14.1
8	114	25	15.1
9	120	60	15.0
10	120	90	16.0

$r_{12} = 0.75$

Figure 3.3a Illustrative artificial example

Simple correlations

	IQ	*Grip strength*	*Age*
	x_1	x_2	v
x_1	1	0.75	0.95
x_2	0.75	1	0.78

$r_{12|V} = 0.03$

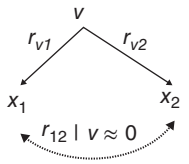
Simple correlation

a) At first, we had found the relationship r_{12}:

$r_{12} = 0.75$

$x_1 \longleftrightarrow x_2$

b) Later, we found a third variable v, which "*explained*" the r_{12}:



"Cleaned" — partial correlation of *IQ* and *strength*:

Age **has "explained" the former** $r_{12} = 0.75$

Compare the $r_{12} \mid v = 0.03$ with the *numerator check*.

$$\textit{Numerator check} = (r_{12} - r_{1v}\, r_{2v})$$
$$= (0.75 - 0.95 \times 0.78) = 0.006$$

We next calculate the scaling factor S:

$$S = \sqrt{(1 - r_{1v}^2)\,(1 - r_{2v}^2)} = \sqrt{(1 - 0.95^2)\,(1 - 0.78^2)}$$
$$\sqrt{(1 - 0.911)(1 - 0.600)} = \sqrt{(0.\,089)0.\,400)} = 0.189$$

We divide the numerator check by the scaling factor to get the partial correlation:

$$r_{12|v} = \frac{(r_{12} - r_{1v} r_{2v})}{S} = \frac{0.006}{0.189} = 0.03$$

Figure 3.3 The "explanatory" role of a third variable v brings the partial correlation to near zero.

different from zero. This may be also assessed by using the approximate *numerator check* of the partial correlation (of the preceding Section §32). Since some statistical methods and models are based on more general kinds of probabilistic relationships than just linear correlation, the more general analog is also known as the *principle of local independence*, which means independence in the sense of conditional probability (which would automatically generate the uncorrelatedness, but not the contrary).

The level to which completeness of "explanation" is approached, in terms of fulfilling the requirement of partial uncorrelatedness/local independence, should be a necessary part of the hypothesis in the corresponding empirical research study. Thus, the "explanatory" task of such a study would make it a *falsifiable hypothesis*—verifiable in a general scientific meaning, and testable in the sense of inferential statistical testing.

3.4 Internal Construct Validity and Latent Factors

Our senior author has mentioned "constructs" throughout within quotes. He is not really sold on how the term is used in the psychometric literature or especially in practice. He is not sure without some digging how to tell whether it offers true instead of illusory explanatory power, predictive power, or power to understand what is going on with persons at different levels of the measurement scale produced by an instrument. In so doing, he leads us into the third main topic of this book, *latent factors and the Latent Factor Model*. This section introduces us to the latent factor model based on what we have learned so far. In later chapters, particularly Chapter 6, he will expand upon it.

Because the Latent Factor Model brings with it powerful evidence for internal and discriminant validity, it offers versatile tools for explanation into the deeper, internal aspects of validation. Well-designed experiments with meaningful external scores can enable deeper understanding as well, through discriminant as well as convergent evidence. Since causal explanation has the greatest scientific merit, Blahuš is giving us insights into how to determine when relationships between variables and items are inherently causal, as an indication that the pairing of a hypothetical construct with its measuring instrument is reaching a high degree of evidence-based validation.

3.4.1 (§34) If the Explanatory Variable is Unobservable, *Construct* a Latent Factor (Construct Validity)

In the examples above, the *explanatory* third variable was accessible to our observation, viz. the overall body height or the age of children. But, what if we believe that such an *explaining* third variable exists but we cannot observe or measure it? For example, we may speculate that a group of several subtests has some *common underlying attribute*, say a common ability, which is displaying itself through the performance of a number of its subtests. It is sometimes said that all the subtests *measure* the same *thing*. Because the hidden *ability* is not observable, we do not know the hidden mysterious scores of individual subjects, the *levels* of their *ability*. Consequently, neither can we compute the correlations of the *ability* with the subtests, nor the mutual partial correlations between subtests after the *influence* of that *ability* on the subtests' intercorrelations has been partialed out.

The psychometric strategy is to construct new tests or locate promising existing tests that seem likely to be assessing the same invisible construct. The next step is to use factor analysis to determine whether or not the scores fall on a single dimension. The third step is to test whether they pass the test of the *principle of local uncorrelatedness*. This will be indicated by the partial correlations between all pairs of indicators being zero or near zero after the common factor (variable *f*) hypothesized to cause the scores on the indicators has been partialed out. If you succeed in finding such numbers, it is reasonable to suppose that this common factor variable *f* is a numeric LCF representing the "ability" hypothesized to have a causal impact upon the indicators.

3.4.1.1 Latent Variables and Local Uncorrelatedness

The following example is a simple but still reasonably authentic illustration of a procedure well known as *modeling with latent variables* (see Chapter 6, especially the [LCF] model in §52).

A computer-based factor analytic procedure is used, which seeks the factor loading values that bring the partial correlations as close to zero as mathematically possible. It is not unusual to fail to get such results. It is an indication that your hypothesis that the tests jointly measure the same "thing" is wrong. In addition, at the first stage of latent variable modeling, we usually do not focus on the latent scores of subjects. Instead, we are interested in the correlations between the latent factor $\boldsymbol{f}$ and indicator variables x_j. These correlations are referred to as factor loadings and also as "latent validities." We should be particularly attentive to the partial inter-correlations between all the possible pairs of the indicator variables after the artificial $\boldsymbol{f}$ has been partialed out. If they are very close to zero, we may consider it as evidence supporting our preliminary hypothesis that the subtests, the indicators "measure" a hidden common attribute, say an "ability," hopefully modeled by the latent factor $\boldsymbol{f}$.

The sketched principles of latent variable modeling will now be illustrated with the example of the CERAD battery with real data gathered in the Utah study of Alzheimer's disease patients. As it was mentioned earlier, the CERAD battery is supposed to cover three "abilities," one of them being "verbal ability." We have already worked with one of the subtests of it, namely the V-Fluency test. Now we will add three further subtests to complete a content-homogeneous four-test battery focused on only one concept—verbal ability. The mutual internal validities (correlations) among the four subtests are in part (a) of Table 3.2. Part (b) gives the factor loadings of each of the four variables on the latent factor $\boldsymbol{f}$. In part (c) of Table 3.2, we find the residual partial correlations after the product matrix $\boldsymbol{f}\boldsymbol{f}'$ has been subtracted from $\mathbf{R}$. They are quite close to zero, indicating that the original manifest correlations in (a) have been "explained" by the mathematically constructed factor $\boldsymbol{f}$ reasonably well. This finding supports our hypothesis that performances on all four subtests may be recognized as manifestations of the same hypothetical attribute, and from the behavioral content of all the subtests, we assume that this is "verbal ability."

Table 3.2 Validity of subtests to the mathematically constructed scores of the latent variable *f*.

	(a)				**(b)**	**(c)**			
	***Internal manifest validity* between the four tests of "verbal ability": correlation matrix**				***Latent validities*: Correlations of the tests with the latent variable *f***	***Residuals* of the correlations after the "influence" of *f* was partialed out**			
	$R = [r_{jk}]$				$[r_{jf}]$	$R_{RES} = [r_{RES}]$			
x_1 Verbal Fluency	x_1	x_2	x_3	x_4	f	x_1	x_2	x_3	x_4
x_2 Boston Naming	1	.51	.52	.43	.70		.02	.04	−.07
x_3 Control Oral Word	.51	1	.40	.54	.70	.02		−.09	.03
x_4 Word List	.52	.40	1	.50	.69	.04	−.09		.01
	.43	.54	.50	1	.71	−.07	.03	.01	

3.4.1.2 A Latent Common Factor Model

The presented example is a little simplistic, but still a good representative, illustration of one of the models for latent variables, the LCF model. This will be described in Chapter 6 in detail (§53). After presenting this numerical example, we should ask what such a procedure means from the point of view of scientific acceptability. Any scientifically-oriented behavioral researcher should be suspicious about such a straightforward simplistic sequence of actions: Just let the computer produce some mysterious numbers, call them latent scores of a fictive ***f***, interpret the artificial ***f*** as a psychological concept, call it "verbal ability," and claim that you have explained the observed data by discovering an "ability." What are the critical moments from the point of view of philosophy and methodology of science in it? To answer the question, several aspects of methodology have to be outlined first:

1) The first one deals with the topic of *scientific explanation*, especially as it has been presented in the former sections on *causal explanation*. It is related to the problem of whether we may accept partial correlation as a kind of explanation.
2) The second one is connected with the *problem of correspondence* between the attributes that are observable by diagnostic methods of empirical research (such as tests) and the *theoretical concepts* of a scientific theory of measurement (such as "ability").
3) The third aspect deals with *representation conditions*, or the level of "measurement" of the latent variable ***f***. These are implied by the researcher's choice of a specific latent-variable model according to his/her theory and hypothesis about the conceptual nature of ***f*** in the investigated area. In the illustrative example in Figure 3.2, the latent scores were quantitative values on a metric-interval scale, which probably correspond to a concept of "ability" and its "grades" or levels. A detailed explanation of these aspects will also be presented in Chapter 6 (§57). First, however, it is necessary to clarify the methodological background of these rather formalistic procedures.

3.4.2 (§35) What are the Behavioral "Constructs" Behind the Scenes? The Theoretical Concepts of Scientific Theory

One of the basic problems of the philosophy of science deals with the relationship between the *empirical* and *theoretical layers* of knowledge—it is between that which is directly observable and that which is not. Various philosophical and methodological approaches differ on this. In addition, the various conceptions also distinguish not only two but *several different layers* between the directly observable and the only hypothetical or even speculative layer. Then the methodological paradigm of *correspondence* between observable and unobservable may represent a rather complicated system of transition from the attributes that are directly empirically observable/measurable to the theoretical concepts. Some of the concepts can have a character of so-called pure dispositional attributes whose nature places them not too far from the "observational floor." Some may be more theoretical. Others may be of a very hypothetical or even temporarily speculative nature. Some social scientists and psychologists like to use the term "hypothetical constructs," although it is clear that many are not literally only hypothetical. As an example, although "spatial short-term memory" is definitely an instance of a concept that is not directly measurable, it is still not as abstract as the concept of "well-being." For the sake of simplicity and the purpose of using latent-variable modeling, let us stay with just two levels of attributes for now.

- *Empirical and observable directly*. This could consist of a measured performance in any of a number of cognitive performance tasks such as the symbol-digit substitution task, the "trail-making

tests," various sensory detection tasks in psychophysical research, reaction time measurements in solving puzzles, or even expressing a certain level of agreement with statements in an attitude questionnaire. To be recognized as empirical research variables in the sense of Section §7, these are typically represented by standardized methods of diagnostic research. They are *concrete*, *partial*, and *specific* in nature, where concrete means *operationalized* in the research inquiry phase. That is, they are accessible through practical standardized research for directly obtaining scientifically recognizable data (see Section §7 of chapter 1).

- *Theoretical and unobservable by direct means*. Although there is no direct way to measure them, they may be inferred indirectly. This could include many more abstract and high-level abilities like verbal dexterity, charismatic influence, intelligence, confidence, depression, optimism, well-being, quality of life, entertaining, persuasive, ability, etc. The character of the theoretical concepts is *abstract*, *synthetic*, and *generic*. The procedure of creating, constructing, or identifying theoretical concepts on the basis of empirical observation deals with a well-known methodological problem—*theoretical concept formation*. The formation process is more or less based on a theory or paradigm of the *empirical-theoretical correspondence*. It is an intuitive and heuristic method of concept formation in the behavioral and social sciences.

3.4.2.1 Theoretical Concept Formation and Latent Variable Modeling

We could point to *latent variable modeling* as a clear example of this kind of *theoretical concept formation*. It is essentially based upon the process shown in Table 3.2. The fundamental explanation of this process is within the scope of this book and will be presented in detail in Chapter 6. Some of the specific methodological terminology will be introduced here. The standardized directly observable research variables are usually referred to as *empirical indicators* of the theoretical concepts, which are the "constructs" we are indirectly trying to measure. The directly observed measures are referred to as "indicators" because we hypothesize that there is a causal relationship between them and the theoretical concept we are trying to measure. The indicators are as they are because of the causal relationship between the theoretical construct and each of them. In our example in Section §34, the indicators were represented by the four CERAD subtests. The indicators share a common theoretical status as ostensibly caused by the trait referred to as "verbal ability."

In the behavioral and social sciences, there are various specific names for theoretical concepts. Besides the already mentioned examples, we find terms like "continua," "dimensions," "subscales," "facets," "aspects," "abilities," and "traits." The researchers who prefer speaking about "constructs" usually do not know that this term comes from the era of an extreme version of positivism nearly a century ago. In that approach, concepts of a theory were supposed to be hypothetically and freely invented or "constructed" by a scientist according to his/her opinion in advance, followed by gathering empirical data for support. From this point of view, the term "construct" is something of a historical relic of the era of logical positivism. It is especially interesting to hear this word frequently used by those who, at the same time, promote a "postmodern approach" and strictly criticize all quantitative methods as positivistic. Presumably, the contemporary practice of the frequent use of the term "construct" is not based on this outdated philosophy but serves merely as a nickname in research jargon due to its brevity.

3.4.2.2 Latent Variable Modeling and the Principle of Partial-Local Uncorrelatedness

If we can demonstrate that when we subtract out from the correlation matrix of indicators all of the predictive influence that is due to $\boldsymbol{f}$, the theoretical construct, and there is then "local independence" in the remaining residual correlation matrix (zero or near zero correlations), we have shown

that our theoretical construct was the common cause of the correlations among the indicators, and we, therefore, have scientifically verified that ***f*** is a latent variable.

This principle of partial-local uncorrelatedness, or as it is also called "local independence," is the essence of the demonstration of true "scientific explanation," by showing that an active change in x produces a concomitant change in y. This result of zero or near-zero correlations among the indicators, after we subtract out everything in the correlation matrix of indicators that is due to our calculated theoretical construct, is analogous to the partial correlation coefficient demonstration shown in Sections §32 and §33. Regardless of whether such a state of affairs as absolute local independence may ever be completely reached, the struggle for it is one of the engines of scientific inquiry.

A part of the idealistic final state of developed theory would be to offer a complete explanation that would not leave any part of an observed phenomenon and its relationships unexplained. It could be considered to be one of several possible methodological aspects of theory building: The principle of *local independence* statistically represents the *methodologically* desirable postulate of *completeness of scientific explanation*. Of course, this is a completeness related to the stage of development of the theory under construction. By common sense, if we are trying to "explain" correlations by constructing latent variables as *constructs* in a formalistic (mathematical) process we must have near zero relatedness after removing the effects of ***f***. We would not be satisfied with "explaining" only a fragment of the correlations, such as leaving thirty or forty percent of the "unexplained" residual correlations as negligible.

This postulate of completeness of explanation holds in general as an important aspect of theory building. Since a genuine *causal theoretical explanation* of empirical findings is generally accepted as one of the main aims of any empirical science, theoretical concepts established by causal evidence are important building blocks for scientific advancement. The evidence for interpreting the theoretical concepts as *explanatory concepts* should be understood as part of the more general problem of *building causally explanatory theory*, including the possibilities of its falsification or its confirmative support. The questions connected with scientific explanation are necessarily related to the general problems of causality.

3.4.3 (§36) Causal Effects vs. Independent/Dependent Variables: A Systems Theory View

Very often, the classification of different types of research variables and their relationships is presented in a confusing way. To keep various notions clear, the S ⇒ R Stimulus-Reaction analogy and a systems theory paradigm of Input-Output variables are helpful tools. The place of variables in the independent-dependent tandem is determined by the *direction of estimation* of values of one type of variables given that the values of the other type of variables serve as a source of information. A typical case is the scheme of *predictive validity* as outlined in Section §29. The predictor variables, such as admission tests, are in the position of *independent* variables, from which the values of a *dependent* variable, say a criterion of academic performance, are to be estimated.

A characteristic sign of the independent-dependent type of dichotomy consists in the possibility of reversion of the two roles of the variables. Then the independent-dependent reversibility and symmetry can be represented graphically by $x \rightarrow y$ as well as $y \rightarrow x$, where $\rightarrow$ means *estimation*. The mutually- symmetric, bi-directional possibility of estimation applies also to common sense in using the two regression equations interchangeably. Symmetry is not only a formal mathematical exchangeability, but it also may be meaningful from common sense. For example, you might be interested in reducing the number of subtests, n, in a battery. Multiple regression and correlation can be used to assess the internal composite validity of tests to an nth subtest. If its scores can be

estimated with a sufficiently small error, the subtest may be omitted. But, you may also ask whether the composite of the subtests cannot be estimated from the *n*th subtest alone, say if you would like to reduce the battery to just one subtest, the best one. As well, you may reverse the former predictive estimation and ask how well you could reconstruct the admission scores by backward estimation of them from the results of academic performance assessment already completed.

3.4.3.1 Causal Variables and Effect Variables

In contrast to this symmetric exchangeability of independent and dependent variables, the role of *causal variables* and *effect variables* is asymmetric and irreversible, at least in general, and particularly in the human sciences. Several kinds of asymmetry are among the fundamental conditions of a causal relation: Asymmetry in *time* makes it impossible to reverse the direction of the time arrow, perhaps with the exception of science fiction or fairy tales. Asymmetry in the *direction of producing* effects means that the *effect variable* cannot work as the *causal variable*. Alcohol in one's blood may cause a subject's reaction time to be longer, but reacting slowly would not produce alcohol in the bloodstream; the rain may cause you to put up your umbrella—but not the opposite.

The dichotomy of causal-effect variables may be represented by the input-output variables of a system since it closely resembles the S ⇒ R behavioral paradigm. The simplest case would be Watson's or Pavlov's black-box system. Note that the active component of a causal variable x is its capacity for creating *change* as does the stimulus in the Pavlovian S ⇒ R behavioral paradigm. This is shown in Figure 3.4. A beginning student of animal psychology might wish to begin his studies by investigating the effects of saccharin on the food intake of lab mice with the hypothesis that the amount of saccharin in a mouse diet of oat mash will be linearly related to the amount of consumption. The x_1, x_2, and x_3 in Figure 3.4 could represent the oat mash at three levels, with deltas Δx_1, Δx_2, and Δx_3 representing the increase in saccharin for each of the conditions over the amount of saccharin in the preceding condition. The pattern of corresponding levels of consumption y_1, y_2, y_3, and y_4 as a result of the amount of saccharin in the oat mash could be used to demonstrate that saccharin increases consumption. Also one could, by varying the deltas in x, test whether the effects on consumption y are linear or curvilinear, whether there are causative effects of changes in the pattern of the deltas in x on the deltas for y, etc.

A *causal* variable is always a special case, a restricted case, of an *independent* variable, and the *effect* variable is always a special case, a restricted case, of a *dependent* variable. Each member of the *cause-effect* tandem has properties that the members of the *independent-dependent* tandem do not, such as the asymmetry in *time* and *direction* that only the causal tandem has, as mentioned in the preceding discussion. We can estimate and derive the values of the effect variable from the knowledge of the manipulated values of the causal variable but the reverse direction from y to x does not hold.

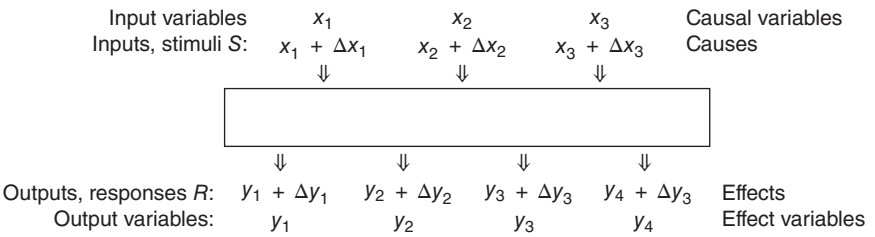


Figure 3.4 Analogy of cause-effect, input-output, and stimulus-response variables.

3.4.3.2 Descriptive Associations and Causal Relationships

Unfortunately, too often the temptation to interpret the correlation of just a descriptive type as if it were a causal influence has misled some authors to use confusing terminology. Thus, some are interchanging the notion of "independent variable" in confusion with "causal variable," or "explanatory variable," or even with "a factor" as a would-like-to-be "experimental factor." One can appreciate this distinction by considering an amusing example of the latter when listening to a doctor's statement:

> Your age and gender are important risk factors for a heart attack.

So, doctor, how are you going to manipulate these "factors"? In addition, some other language tricks are in favor of those who either do not know about the difference or do not care about the ethical imperative in science. There exists a bad habit of calling the independent variable "determinants," or saying that a dependent variable is a "function of ... (something)." Both of these mean *causality* in disguise. In addition, some pure mathematicians and theoretically oriented statisticians in interpreting a regression function, might interpret the raw score regression line and its slope coefficient b by saying that it indicates a "change of a certain amount in y when x changes by one unit." They forget the fundamental nonmathematical but *methodological difference* regarding from *whence that* "*change*" *has arisen*. It is one thing if you "change" the value of body weight, as the independent variable x, just by asking another individual in the sample to step on the scales, while it is a quite different thing to have change in the body weight of one particular individual measured at two separate times as a result of a dieting and exercise regime as part of that person's wellness program.

Therefore, one must *not* misinterpret two substantially different kinds of "change" that represent two different *methodological types of variability*. One is *inter-individual spontaneous* variability under natural conditions, which is, in addition, occurring simultaneously as a synchronic symmetrical association. The other is *intra-individual manipulated* variability due to an experimental treatment produced with a time delay, as a diachronic asymmetric causal response.

3.4.3.3 Systems Flowchart vs. Path Diagram

Let us consider a simple case with just one output variable y, that is, one *effect* variable. Then the black-box system paradigm on the left of Figure 3.5, known as the flowchart, can be considered a special case of the independent-dependent variables graph shown on the right—the so-called *path diagram*. Each of the two methods of graphical representation may be very well represented by a so-called *directed graph*, or "*digraph*," employed by the formal and mathematical discipline known as

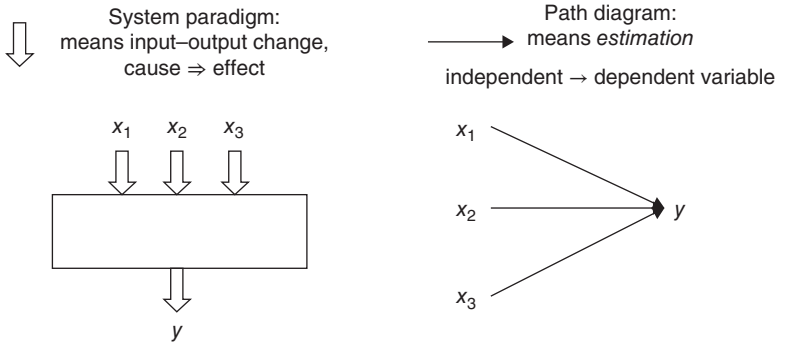


Figure 3.5 Comparison of the system flowchart vs. the path diagram.

graph theory. However, we use the systems theory paradigm and path diagram to express two very different kinds of change. The behavioristic Stimulus | Reaction active causal change is shown in System Paradigm on the left, the causal special case. In contrast, the less commital path diagram on the right, the more general case, merely shows that x_1, x_2, and x_3 are all related to y, but we do not know whether the *x*s or the *y*s have a causal relationship or which way the connection could go, only that $x \Pi y$ as presented on the right side of Figure 3.5.

The former, the causal relationship of the system paradigm is represented in the figure by | in the text and by a system box with a vertical *top-to-bottom* orientation. The latter, the path diagram on the right side, is expressed by a *left* Π *right* orientation of the "path of estimation." Contemporary study in the theory of directed graphs as related to causality and path analysis has created new and useful knowledge, mostly due to Judea Pearl (2000) in his famous *Causality* monograph. In light of this new approach, the $x \Pi y$ path is considered the probabilistic dependence of y on x, which enables the estimation of y by a *probabilistic operator* Π, say a regression estimate. However, the causal change of the value of a variable and its active influence is represented by Pearl's (2000) "*do*-operator" $x \mid y$, and his corresponding "*do*-calculus," which deals with active changes, not just with passive/descriptive estimation through the traditional techniques of probability theory and regression.

3.4.4 (§37) Theoretical Concepts as Active Subsystems Hidden in the S ⇒ R Black Box: Causal and Propensity Interpretation of Ability

The behaviorist S ⇒ R paradigm faced a difficult question: How can the stimuli on input be transferred to the output in order to manifest themselves as responses? How can they be interconnected inside the organism? Or, to speak in contemporary systems theory language, how are they interconnected inside the black box? Even in the contemporary philosophy of science, some interesting debates are held on the problem; some authors bring arguments that such a connection is necessary (e.g. Dowe, 2000; Salmon, 1998), while others are trying to prove the opposite (esp. Schaffer, 2000). Nevertheless, in the times of the behaviorist nineteen-thirties, the idea of *S–R* connection was quite appealing. Among the answers, there was Tolman's (1932) logically derived theory, which hypothetically inserted hidden "intervening variables" inside the system. It was assumed that they were transmitting the internal flow from x to y. Speaking in today's language of systems theory, the intervening variables worked as transfer functions transforming x into y. Tolman's idea is conceptually fruitful since it stresses the *processing character* of the theoretical concepts ("constructs") as mediating *active* entities. In his modification, the S → R scheme got the form S → O → R, with O for organism intermediating the connection. Thus, this causal interpretation of "constructs" becomes one of several possible approaches to their conceptualization.

There are at least three *interpretations of the empirical-theoretical correspondence*, three methodological types of *associating functions*, which may be assumed to interconnect empirical indicators with their theoretico-conceptual variables. Or, seen from the other direction, these represent three ways of the *manifestation of a theoretical concept* through the indicators.

1) *Causal interpretation*. A theoretical concept ("construct") may be assumed to work as a certain type of *common cause* of its manifestations in the indicators whose observations are supposed to be the effects. Namely, it is an instance of a type of cause that is sometimes called an *immediately preceding cause*, i.e. preceding in the causal chain of causes and effects leading to the final effect. For instance, an appropriate "amount" or level of verbal ability produces successful performance in a practical task, in which that ability is necessary. But the first, starting or *triggering cause* is a stimulus that "excites" the potential "ability," which, consequently, produces an observable response. In the sense of Section §31, the *counterfactual condition of causality* can

be illustrated here very well: If there had not been "enough ability" the observable response would not have been produced. (See Tolman's intervening variables.) This also represents the theoretical concept as an *intermediating* and a *potential cause* in the causal chain: It does not produce the effect if there is no triggering cause. The trigger may become a stimulus in the form of exposing the subject to a specific behavioral situation, or a test instruction, or a command given by the experimenter, or a question asked in an interview, etc. Its system scheme is pictured in the left part of Figure 3.6.

Some researchers criticize this approach perhaps on the same ground that it is too similar to searching for "a ghost in the machine" (in the sense of Ryle, 1949). Namely, such critique gets a special form by those who use models with latent variables for identification of "constructs" and conceptualize the modeling only as a kind of measurement, however in the broader sense. In such an interpretation, the models are just more general analogs of the Gauss errors-of-measurement model, and of the models used in engineering, technology, or geodetic measurements. This interpretation is described below as the third variant. However, it seems quite natural that theories influenced by behaviorism more or less explicitly assume that the theoretical concepts play the role of an *intermediating causal potential* in the Stimulus ⌋ Response chain. Again, let's recall the role of Tolman's intervening variables.

2) *Propensity toward a behavior*. Another conception assumes that some of the theoretical concepts in psychology may be interpreted as the inclination to a certain behavior, as *tendencies* or *habits to behave* in a certain pre-programmed way with rather high predictability and probability. Then this *probability of empirical manifestation* associated with various levels of the theoretical concept is regarded as a *propensity function*. A subject with a high level of an unobservable trait of personality, say "Extraversion," will display corresponding characteristic behavior in different observable concrete ways in the indicators. For example, he/she will answer the questions of a standardized questionnaire in a certain way. If this conception assumes that the behavior is displayed only as a response to certain stimulating situations, then it is not too different from the preceding case of Stimulus ⌋ Reaction and causal interpretation. A classical standpoint may be illustrated by the historical definition of Allport (1931):

> ... to know that an individual has a trait it is necessary to have evidence of repeated reactions, which, though not necessarily constant in type, seem nonetheless to be consistently a *function of the same underlying determinant*.

Here the formulation in italics seems to be a cause-effect relation in disguise.

On the other hand, it is not quite clear whether this conception also allows for the case where the observable behavior is produced only spontaneously without any specific stimulus—then the system scheme in Figure 3.6 would have no input arrows. A naive illustration could be that

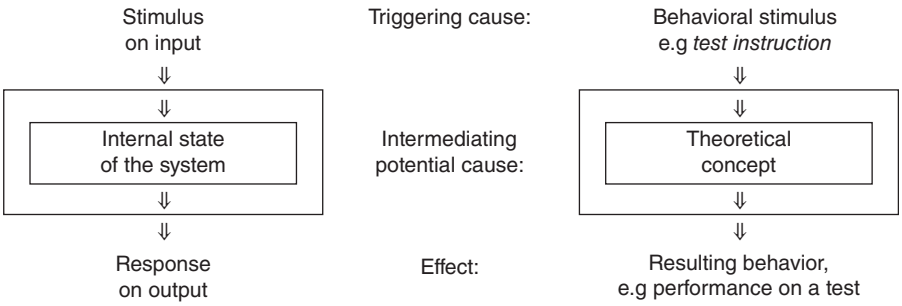


Figure 3.6 Theoretical concepts—that is, "abilities"—as analogs of the internal state of a system.

the individual with a high level of "verbal ability" would be expected to display his/her ability spontaneously, being talkative. Rather, a more realistic case is where we carry out a purely descriptive study not being interested in the specific inputs that produce the observable behavior in the indicators. In such a case, the propensity would produce *manifestations that are typical on average* within the frame of our pre-specified representative sample of subjects under study. Then we face the typical situation of so-called "correlational research": We are able to find only *observational associations between the indicators*. This fact is related to two possible aspects of the propensity conception.

2a. *Inductively synthesized concepts* are hypotheses about the hidden internal structure of the black-box system made after the fact. If several output indicators, which are conceptually *homogeneous by the content of the behavior*, create a highly correlated *cluster*, we may hypothesize that our observations of them were manifestations of their *common concept*, a hidden element in the system. In the system paradigm below, see the example of the two unobservables that are hypothesized to underlie performances in the CERAD battery, namely verbal ability and executive ability.

2b. Deductively derived manifestations of observable behavior appear as if the kind and intensity of the not-yet-manifested habitual tendency were known in advance. Given that we knew the true general "level" of "Extraversion" in an individual we could expect—*estimate* or *predict*—his/her specific behavior in usual situations. See the causal/systemic vs. path diagram of the ability example in Figure 3.5. In the case of the propensity conception, i.e. while we are observing the *behavior as typically produced by the internal habitual tendencies* under usual or standard stimuli from the environment, the system paradigm and path diagram become the same as shown in Figure 3.6. This kind of interpretation somewhat resembles the already mentioned problem with the habit of using the term "construct." The fact that the hypotheses about theoretical concepts as abilities etc., were *created in advance* does not mean that they were "constructed" purely artificially and arbitrarily. Again, we would rather believe that the term "construct" is habitually used just for brevity and as research jargon.

3) *Measurement Approach*. A third interpretation, maybe not the last one,[8] is based on the idea that the substantive behavioral attributes of a system are accessible only indirectly—through *observation channels* that are influenced by noise. If the "true" attributes are of a quantitative nature, we may specialize the observation process as a kind of measurement and treat the *noise* as "errors of measurement." Thus, this interpretation can be called the "measurement approach" to the conceptualization of theoretical concepts, e.g. abilities. Speaking the language of systems theory, the noise is assumed to be so-called *white noise*, which means that the errors are considered purely *random*. Among other things, it means that the *errors are independent* of the "true" level of the indirectly observed attribute.[9]

8 Here is another similar measurement approach taken by Mari, Wilson, & Maul (2021) *Measurement across the Sciences*. They however propose a measurement approach that includes a causal interpretation as in the Borsboom definition of validity discussed in 3.0. Theirs also may not be the last one, but it could become the most influential in its time, as it has developed new definitions of the many terms in Metrology as proposed updates to the VIM so future versions of the Vocabulary International of Metrology may include them. This leads to a solution for Petr's frustration with inconsistently used terminology.

9 *Editorial note:* A Measurement Approach is taken in the entire book by Mari, Wilson, & Maul (2021), but it incorporates the causal model as well. They also discuss the independence of errors on pp. 166–167, Section 6.3.7. These specific objectivity assumptions of the Rasch model also lead to a quantitative scale (interval) for Reading Comprehension Ability. Thus progress over the passage of time has seconded Petr's third interpretation, and incorporates his recognition of the independence of errors, and his aspirations for quantitative latent abilities.

The two former interpretations—causal and propensity ones—had considered theoretical concepts in the role of internal *elements/subsystems* that function and produce observable responses in the system's output. The present approach seems not to care about functioning, about the mechanism, or about "how the thing works." Instead, it reduces the problem just to "diagnosing" the "genuine" attributes and their "true" values. However, if we still try to combine this measurement approach with the former propensity interpretation of "ability," then "ability" could be taken as just a "potential stockpile" whose "true level" is to be estimated at a given moment through a channel of observation. The channels for observing "ability" are represented by a cluster of content-homogeneous subtests in the role of empirical indicators of the theoretic-conceptual attribute. This seems to be somewhat similar to the method of triangulation in geodetic measurement: The value of a certain variable is estimated from several measurements of the other variables. In general, the process consists of two stages of representation and formalization that are summarized in Table 3.3:

The first formalization stage—establishing scientific variables combines levels (a), (b), and (c) in the four main progressing rows of the table. It deals with a transition from the level of (a) *pre-scientific attributes* to the level where these are formed into (b) *quasi-scientific attributes*, and finally, to the level of (c) genuine *scientific variables*. On level (b), there may exist in practice many *clinical and educational practical diagnostic procedures* that work well for current purposes and are accompanied by a theoretical background consisting of theoretical "constructs" that are often speculative and without a clear status as genuine, though *not-yet formalized scientific variables* on level (c).

The second formalization stage consists of a transition from the unformalized level (c) to the formalized or mathematized level (d) in Table 3.3.

3.4.5 (§38) Theoretical Concepts as *Attributes* Under *Indirect Observation*, with Their Indirect Measurement Through the Manifest-latent Link

3.4.5.1 Manifest and Latent Variables

The whole process results in forming the *manifest* and *latent* variables in the final form of fully formalized/mathematized variables after passing the process of transition indicated in Table 3.3. The transition from the substantive primary pre-scientific level (a) to the still substantive, though not necessarily formalized, minimum-scientific level (b) is related to going from gnoseological aspects of the *philosophy of science* to a general scientific *methodology*. This would recognize the dichotomy of the *empirical* "diagnostic methods" vs. *theoretical concepts* ("constructs"). Further, these two are treated as more specified forms of *scientific variables*, namely the *empirical research variables*, in the role of *indicator* variables vs. *theoretic-conceptual variables*.

On level (c), the two are considered in their more rigorous scientific forms: An empirical indicator variable becomes a standardized research method for the scientifically based observation/measurement of empirical data whose coding, however, need not fulfill the complete conditions for their mathematical treatment. Such a theoretic-conceptual variable represents its preceding theoretic concept, but now in a more rigorous scientific form of the designated scientific theory. Finally, scientific variables are formalized/mathematized on level (d), and become manifest and/or latent variables, usually in a symbolized form of abstract mathematical objects, typically the stochastic-random variables of probability theory and mathematical statistics.

Petr Blahuš[10] is an enemy of shoddy but expedient practices, often used and even published, that do not lead to scientific value or the best professional service. In this final work, he tries to show his

10 This paragraph and the next two, then §38b are 2023 commentary on Table 3.3 and how it may be instantiated by a Measurement Approach following the 2021 book on *Measurement Across the Sciences*.

Table 3.3 Levels of transition from pre-scientific attributes via unformalized scientific ones to their representation by formalized/mathematized variables.

Observable to be explained (explanandum)		Unobservable explaining (explanans)	
Pre-scientific (a)	*Real-world* observable *attribute* • sometimes partly formalized or even quantified • –not standardized	*Speculative* unobservable *attribute* • categories of the attribute unclear • quantitative vs. qualitative unclear	
Quasi-scientific (b)	***Research "diagnostics"*** with so-called. "***standardized"*** instructions, conditions	***Theoretical concept*** (***"construct"***) with • tentative categories of the "construct" • only hypothetical structure, if any • not necessarily formalized (yet), maybe with a crude quantification (e.g. sum of scores of items of a subscale)	
Scientific variable (c)	• A general point of view: ***Empirical research variable*** • The "Concept-indication" view: ***Empirical indicator variable*** – standardized – categories logically consistent – formalized/quantified according to measurement representation conditions	**Rules of correspondence**	***Theoretico-conceptual variable*** • categories in a logically consistent structure • conceptualized usually (!) as a scalable variable or even metric quantity
Formalized/ mathematicized scientific variable (d)	***Manifest random variable***, quantity in the idealistic case	**Indirect measurement: associating function**	***Latent random variable***, (*a quantity*, in most latent models) Named by different models as: • latent "common factor" • latent "trait"/latent "ability" • latent "class" (qualitative, nominal)

readers a path from the initial vagueness at the beginning of exploration of a domain having potentially measurable attributes ***(or properties, as in VIM3)*** to a final truly scientific set of variables. Table 3.3 summarizes at the first level his views of how to get to good measurement with a mathematical representation that traces back to the meaning of what is real to be studied. Without detailed steps of design, it is a high-level model of progressive levels of development of valid measurement instruments in the human sciences, in the form of a table progressing from early exploration top down, culminating with scientific variable(s) that might actually contribute to the science of a field or to valuable applied work.

This chapter cites many references published after his death. It begins by bringing the concept of validity up to date. Now we have more. There are downsides that his manuscript was not published for 15 years, and then only in part. The upside is that the successor authors have been able to see where ideas he understood and did not have time to complete have progressed through later research. For example, we explain in Chapter 8 where his partially developed ideas of Ostensive Definition and Demonstration could still go.

In Sections §37 and §38a, he has communicated his concept of a disciplined, principled process going from vague ideas and weak initial scales through several stages of disciplined reasoning and data collection. Table 3.3 lays out this process. It could well be described as the outline of *an intentional process of measurement instrument design and development*. Despite the fact that he did not mention the work of Mislevy and numerous associates in the development of ECD, nevertheless, in his own language and understanding, he tries simultaneously to discount careless and unscientific, but common, practices and replace them with principled, logical, and empirical processes. Table 3.3 is his Principled Instrument Design overview. He expands on elements of it in Chapters 4, 5, and 6 where we will comment and reference the best we are aware of regarding the latest expression in current literature where his ideas have found support in the work of others, including some of our own. We will also show where we think he needed a course-correction in his honest searching and thinking toward better alternatives to accomplish what he was seeking.

3.4.5.2 Reading Comprehension Ability as a Formalizable Theoretical Construct

We have replaced Petr's discussion of verbal ability with more recent similar work that speaks to the same issues and uses Reading Comprehension Ability (RCA). Summarizing a few things he wanted to cover—Petr spends a paragraph going from level a to level c fairly quickly with the verbal ability example. There is a long history of verbal ability and many instruments to measure the construct. Thus, level a) is long ago, and Table 3.3 includes at level b "crude quantification, e.g. sum of scores of items for a scale or subscale." Otherwise, he sees verbal ability at his level c of scientific variable or higher. Now we will turn to Mari et al. (2021), who picked a dimension of the multidimensional construct, verbal ability as their example of a measurement approach to the problem of linking from empirical observables to their interpretation and use in what they call "the information world." Verbal ability includes oral reading out loud and listening (discussed in Chapter 5), not just knowing words and understanding them through reading. Thus, RCA is closer to a single dimension, but word recognition and understanding may split off as a separate factor from Comprehension.

Mari, Wilson, & Maul (entire book)[11] carried along three main examples: one from an *Extensive Domain*, length, an SI Base unit; one from an *Intensive Domain*, temperature, also an SI base unit; and, as mentioned, RCA. We would call this an *Ostensive Domain*, a term introduced to us by Petr. It is interesting that all three of these instances were treated in detail by Petr over a decade earlier. In addition, he provides an explanation of direct and indirect measurement, and so do MWM (2021, Chapter 7.2), but these are widely divergent. They define the measurement of RCA with a well-designed and calibrated instrument as Direct, while Petr uses the term Indirect. Petr puts verbal ability in the unobserved (but explanatory) column, thus indirectly measured. MWM defines three measurement processes that make a measurement direct, and all three measurement processes could be built into an instrument or system with today's technology. By detailed development of their definition of measurement and its direct and indirect methods, they dispel years of ambiguity and circularity from writers out of a wide variety of disciplines to create their definitions. Putting their definition of direct measurement into an instrument makes it rather obvious that it meets a clear and highly useful definition of direct measurement. It is obvious to those familiar with many varieties of online testing technology that this makes sense.

Before explaining how an instrument can handle RCA, let us turn first to the thermometer as it is easier to explain its causality and how it implements direct measurement. As a backstory, imagine it is desired for scientific purposes to measure the temperature of an object in the SI base unit called

11 For short-hand communication, we will abbreviate frequent references to this book as MWM (2021 {followed by page or chapter section}).

Kelvin, which has an absolute zero and is known as thermodynamic temperature. Only an old glass Celsius thermometer was available at the lab at the time, but for the purpose, its accuracy was close enough. These two scales use the same unit but have different 0s. The object to be measured is brought in contact with the Celsius thermometer. At this time, the alcohol (previously shaken down from prior readings) expands. The measureand, heat, causes the sensor, alcohol, to change its size and that expands the visible line up the glass tube. This causal occurrence is called *Transduction*. It is the first of three processes in the black box that goes from the observable empirical world to the *information world* in the MWM definition of direct measurement. Its causal nature is why we say that the measurement approach of MWM encompasses both the #3 measurement and the #1 causal approaches to bridging the gap between the empirical and the informational world where, as Petr noted, explanations must be available. It is the first "associating function" Petr sought to bridge between the left and right sides of Table 3.3. Transduction produces the *instrument indication,* the visible level the alcohol reached. The *instrument indication* is not a value of use in the information world.

The second process within the black box, and a second "associating function" Petr was seeking, is called *instrument scale application*. The instrument indicator on the old Celsius scale is etched or cast on the glass. This means that a human observer must read the end of the alcohol, and write down the result in Celsius units. When instruments do not include sensors to read the *instrument indication*, the operator may need to become part of the instrument within the black box and observe and record the value from the Celsius scale. This is an optional process called *"matching"* as part of the *instrument scale application* process. This temperature the observer writes down is the *indication value. Transduction* and *instrument scale application* are the first two necessary parts of the measurement function that links the empirical-observable with the theoretical-unobservable. Any value, including the *indication value,* has bridged across to the informational world.

The third and last process within the MWM black box is required when the local scale, internal to the instrument, is different from the *public scale*. As mentioned, the measurement described above was on the Celsius scale. As Kelvin units were required, there will need to be a further linking function, *Calibration,* to convert the Celsius scale reading to Kelvin, the public scale required in this example. In this example, the observer will also have to provide the *Calibration* function by using the simple linear equation to convert from Celsius to Kelvin. This is not hard because they have the same unit, 1/1 so the observer just adds 273.15 to the Celsius value. This additional Measurement Link will be dealt with in Chapter 4.

Now consider RCA. The instrument could be a Reading Comprehension app on a cell phone that can provide all three functions of Direct Measurement as discussed above in the case of temperature. There need be no requirement for an instrument operator to step into any of the processes of *Transduction*, *Instrument Scale Application*, or *Calibration*. In our example of an app on their computer, or tablet, there may be another app that administers the test online and delivers the vector of scores to the app the student is using. The test administration online system presents a series of questions that each follow a reading selection. Online proctoring is available and many online educational programs use them to reduce cheating. The answers to the questions create a column (*vector* in matrix algebra, see Appendix 3) of 1's and 0's for correct or incorrect. If the internal scale of the instrument is only ordinal, the sum of these 1's is the *indication value*, or ordinal value to be compared amid the scores of other students. (This is at level (b)—in Table 3.3). The student using the app will find that her *indication value* (*or* score) is computed immediately. Assume the app can contain past data and connect with a data base of others taking this test. The collection of the item scores and summing them up is the Transduction event. In this simple level (b) case it is adding the scores up and displaying the results on a norm table as a percentile or other common score. It could be said that this is direct measurement by an instrument that measures at the ordinal level.

Now we move to Measuring RCA on a quantitative scale. To obtain such a scale requires the independence of errors assumption. In MWM (2021, subchapter 6.3.7, pp. 166–167) the application to RCA was discussed. The specific objectivity assumptions of the Rasch model when met lead to a quantitative scale (interval). Specific objectivity requires that both the errors in the student parameters and also errors in the task parameters are uncorrelated. Georg Rasch demonstrated a logit scale (log interval) and a nonlinear model that is now classified as a special instance of item response theory (you may practice it in Computational Guide 3). The parameters of logistic functions must be estimated by an iterative process to obtain both item difficulty and also the ability of persons. Once those parameters are estimated, ideally over thousands of cases, the task parameters and procedures for estimating a person's score directly through an instrument follows. Assuredly this is more complex than as discussed in the case of simple sum correct scores for RCA. The instrument would require work to make the complex details of direct measurement invisible to the student. Computer adaptive testing (discussed in §52c) could be used to minimize the number of questions needed in online testing. The term "direct measurement instrument" in MWM does not require a single handheld instrument. An interactive system over the internet is usually the context for applied implementations.

This completes the explanation of recent work which brings Petr Blahuš's yearnings much closer to reality. If we may characterize his Table 3.3 we would call it his tabular framework for instrument development and evolution toward scientific and pragmatic excellence.

3.4.6 (§39) On Terminology 2: Mishmash Around Constructs[12]

The classification of the levels in Table 3.3 is also motivated by an effort to maintain certain minimum language conventions and avoid confusion due to the mixing of terms that belong to different levels and different branches of science. In many textbooks, an everyday laboratory jargon is mixed with almost journalistic metaphors using terms borrowed partly from psychology, methodology, philosophy, experimental design, statistics, matrix algebra, geometry, and computational methods. Petr was an advocate of the opinion that after some background explication of terms has been agreed, their meanings must not be considered flexible until a new variant of explication is agreed. This view corresponds to Carnap's (1956) term of "explication" as a tentative mutually agreed preliminary definition to eliminate some possible confusing connotations of certain notions (vs. "explanation") as the basic purpose of scientific inquiry. If there are homonyms with the same term having different meanings in different branches of science the researchers working across interdisciplinary boundaries must be warned about a possible misunderstanding, and the original meaning of a given terminology must not be imported into other disciplines with ambiguous meaning.

---

12 The senior author was determined that the field of psychometrics and related domains needs clear and consistently used terminology and he is correct. Sometimes, his efforts to provide terminological clarity are themselves confusing by showing so many variations. In Section §39 he expands on the terminology problem and makes some recommendations. Then in a part of the manuscript we did not print, he offered a long, commented list of other uses of the terms. We have listed his preferred terms and definitions, but deleted the other comments in Section §39 for the following reasons.

1) There is an authoritative dictionary for terms in international metrology. It is the VIM, Vocabulary International of Metrology, maintained and updated by the Joint Committee on Guides to Metrology, (JCGM) of the BIPM (see chapter 8). Additions to the VIM that broaden it to both ordinal and interval variables in the Human Sciences have been proposed by Mari, Wilson, & Maul (2021, appendix A).
2) His manuscript was already far longer than suitable for a single book.

3.4.6.1 Formalized (Mathematical) and Unformalized (Scientific) Variables

Therefore, once the distinction between formalized vs. unformalized levels, as well as between the methodological vs. mathematical/statistical terms has been established, these must not be interchanged or mixed. Therefore, I strongly suggest that the terms of the dichotomy ***manifest*** *variables* vs. ***latent*** *variables* be used only for *formalized* variables, of course including mathematical or statistical notions—in the ideally demanded form of metric quantification as a dichotomy distinguishing two types of *mathematical variable quantities* (usually in connection with latent variable modeling.) On the other hand, the terminology of ***empirical indicator*** *variables*/***theoretico-conceptual*** (scientific) *variables,* should be used for the *unformalized research/scientific* variables.

3.5 Summary

The three big parts of this chapter are Validity, Explanation, and Constructs. We have journeyed through some foundational terms and methods in Section 3.1 and in Appendix 5 that together give us a view of what multiple regression is, and how it applies to the first aspect of validation mentioned—Test-Criterion correlations, or regression predictions of criteria from test scores. This is a fairly simple answer to Maria's question at first. A test used by a therapist must be accompanied by validation evidence that it is related to the issue the patient is having, evidence that it will lead to some desirable outcome that can be observed and assessed on some important criterion of progress in the patient's treatment plan. Much more deeply, had she mastered this chapter, she would have begun to understand that there are a lot of very questionable, poorly defined, and poorly supported "constructs" out there that may have only some correlational evidence toward valuable outcomes. Is a new measure of some supposed construct actually relevant to explaining the condition of a patient, or the progress a patient has made, or what to prescribe next in a treatment plan? We do not want our loved ones who need therapy to be exposed to therapists who are trying to interpret and explain the patient's real situation, equipped only with "so-called constructs" in the upper half of Table 3.3, and on instruments developed to supposedly measure them, that measure it only partially, measure other irrelevant variables, or even things that provide harmful advice to the patient. Validity is indeed the most basic concern in measurement of human conditions and progressive attainments.

On the broad topic of validity, our senior author started at perhaps the most concrete aspect and oldest historically of external validation, the test-criterion relationship, then showed us its weakness—its tenuous connection to causal explanation. Knowing that the reader was not forewarned that the entire concept of validity and validation was far more than this, we provided a discussion of validity and validation, containing recent references dated since our senior author's final illness. Yet he anticipated the central role of causal explanation at the core of the validity concept. How does the invisible ability or personality trait or (often more visibly) physical skill interact with the world? Has a proposed new instrument that claims scientific respect gone through a rigorous instrument design, development, and evaluation process, assessing all the evidence for internal validity? Is there internal and external validation evidence sufficient to claim that the scores from the instrument can be interpreted in a scientific, rather than in a speculative and vague, way?

In this chapter, we have obtained our first look at the fourth big topic of the entire book, the Latent Factor Model. We have been given the hope that latent variables and factors, and the computational and experimental design tools associated with them, can lead us to develop and validate instruments that measure constructs that can make a difference in the lives of people. The chapter ended with great promise. The End Notes in Section §38 put his work in context with the recent

work by Mari et al. (2021) *Measurement across the sciences;* also, Mislevy (2018) and his now well-known work with many collaborators and current users of ECD for assessments. Petr's work summarized in Table 3.3 is an instance of his own concept of a measurement instrument being an intentionally designed process that takes time and research along the way to obtaining a real and valuable measuring instrument. His intuitions and aspirations were in tune with these recent developments.

3.6 Study Questions

3.1 Validity and Validation

Q1 Carefully read the account in the introductory paragraphs of the *validity argument* position (Unified Validity) of Samuel Messick, Michael Kane, and Steven Sirici, and write a brief essay explaining that position. The "unified" in Unified Validity means including everything in the testing context, including "culture fairness," how the scores are used, etc.

Q2 From that same introductory material, write an essay explaining the contrasting *Instrument Validity* position of Borsboom and others. You may wish to also bring into that discussion their emphasis on causal explanation as referred to in questions 12 and 13 that follow, and why they might be impressed with Petr Blahuš's psychometric examples from kinanthropology and the Olympic Games where causes are often discoverable. Source: Borsboom et al. (2004).

3.2 Terms and Tools in External Validity

§26 What is Validity? External, Internal, Concurrent, Predictive, and Practical Utility

Q3 What is criterion validity?

Q4 What is external validity? What is internal validity? They are not so much "varieties of validity." It is better to say they are ways of demonstrating validity. Give examples of each: construct validation, face validation, content validation, concurrent validation, convergent validation, discriminant validation.

§27 Is the Error of Estimating the Criterion Sufficiently Small for Some Test Uses?

Q5 How is variability related to validity? What is composite validity?

§28 Differential Weighting of Subtests: Composite Validity as Multiple Correlation

Q6 How does proper weighting improve composite validity? What statistical methods are effective in maximizing composite validity?

§29 A Criterion Score is Best Estimated by a Particular Composite—Multiple Regression

Q7 Discuss the three methods of constructing composite scores. What are the advantages of each?

Q8 Explain the difference between standardized multiple regression and unstandardized multiple regression.

Q9 What is a multiple correlation coefficient and how does it relate to multiple regression? How does one interpret a multiple correlation coefficient?

3.3 Explanation Requires More Than Correlations

§30 *Scientific Explanation and Prediction*: Multiple Correlation and Regression as Its Statistical Metaphor (i)

Q10 On the basis of your reading of Section §30 and your knowledge of multiple regression from Appendix 5, write an essay regarding how multiple regression is expressive of and related to the fundamental enterprise of psychometrics, that is, to the task of obtaining an estimate of a proper latent variable from an optimal weighting of its indicators.

Q11 In what ways is multiple correlation a metaphor for scientific explanation and prediction? Does it constitute bona fide scientific explanation and prediction? What is it missing?

§31 Correlation vs. Causation: Their Respective Roles in Scientific Explanation

Q12 Write an essay explaining the authors' contention in the third paragraph of Section §35 that the highest level of explanation in science is based upon the establishment of causation, and the related maxim that there is "no proof of causation without manipulation." Why is the demonstration of an "active change in x producing an active change in y" particularly convincing as a causal explanation?

Q13 Write an essay based on the fifth paragraph of Section §35 comparing what the authors refer to as the "two types of correlation." You may wish to contrast the explanatory power of the first type, namely "manipulated increments in the predictor variable and the corresponding increments in the criterion variable produced by those manipulations," with the weaknesses of the second type, "passively observed descriptive correlations."

§32 The "Cleaned Off" Incremental Validity of a Subtest: Partial Correlation

Q14 What is meant by cleaned-off validity? How is it explained by partial correlation?

Q15 Read the subsection under the heading "The dilemma of internal and external validity." Identify what that dilemma is, and write an essay on the "well-known compromising rule," that additional subtests (or indicators) chosen to amplify the original indicator should have a relatively high validity (correlation) to the external criterion but also a relatively low correlation to the formerly chosen indicators.

Q16 In Subsection §36, in the last subsection titled "Calculating partial correlation," where the formula for partial correlation is given, write an essay on the rationale for the formula, especially the numerator where the first term is an observed correlation between two variables and the second term is the product of two correlation coefficients, captures the amount of the covariance (correlation) between the first two that is explained by the third variable. You may wish to discuss how this is reflective of the astounding efficiency of a good latent variable in accounting for nearly all of the magnitudes of all the coefficients in the 4×4 correlation matrix, as shown in Table 3.2 of Subsection 3.3.4.1 constructed by the senior author.

§33 *Scientific Explanation of a Relationship* Between Two *Variables*: Partial Correlation as Its *Statistical Metaphor (ii)*

Q17 Explain "cleaned-off validity" in terms of the example in Figure 3.3 of the $r = 0.75$ correlation between grip strength and intelligence.

Q18 How is partial correlation a metaphor for a true scientific explanation?

3.4 Internal Construct Validity and Latent Factors

§34 If the Explanatory Variable is Unobservable, *Construct* a Latent Factor (Construct Validity)

Q19 Explain Blahuš's demonstration of the "LCF model" as shown in Table 3.2. That is, explain how the latent variable can be created as a single variable maximum likelihood factor that is shown to be causally linked to the indicators through the demonstration of "local independence."

§35 What are the Behavioral "Constructs" Behind the Scenes? Theoretical Concepts in Science

Q20 Explain Blahuš's distinction between "empirical attributes" and "theoretical attributes," and his concept of the "empirical-theoretical" correspondence. What is weak associative measurement? What is the principle of local independence (also referred to on page 43 as "partial-local independence")?

§36 Causal Effects vs. Independent/Dependent Variables: A Systems Theory View

Q21 How does the contrast between systems theory flow charts and correlation/regression path diagrams help clarify the difference between causal explanation, that is, cause/effect relationships, and correlational relationships between independent and dependent variables?

§37 Theoretical Concepts as Active S*ubsystems* Hidden in the $S \Rightarrow R$ Black Box: Causal and Propensity Interpretation of "Ability"

Q22 Summarize the three interpretations given in this section of the "empirical-theoretical" correspondence used for interpreting or explaining a propensity or a construct such as verbal ability or executive ability.

References

Allport, G. W. (1931). What is a trait of personality? *The Journal of Abnormal and Social Psychology*, 25(4), 368–372. https://doi.org/10.1037/h0075406

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1–16.

Angoff, W. H. (1988). Validity: an evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity*, (pp. 19–32). Lawrence Erlbaum Associates, Inc.

Armstrong, J. S. (1975). Tom Swift and his electric regression analysis machine: 1973. *Psychological Reports*, 36, 806.

Blahuš, P. (2000). Measuring and modeling motor abilities as concept formation. In D. Milanovic (ed.), *Kinesiology for the twentieth century, Zagreb*, University of Zagreb Press, (Croatia), 43–50.

Borsboom, D. (2005). *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge University Press. ISBN 978-0-521-84463-5.

Borsboom, D., Cramer, A. O. J., Kievit R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications*. Information Age Publishing.

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement*, 50(1), 99–104.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.

Box, G., & Draper, N. (1987). *Empirical model building and response surfaces*. New York: Wiley.

Braun, H. I., & Wainer, H. (Eds.). (1988). *Test Validity: Papers of a Conference Entitled "Test Validity for the 1990s and Beyond", Held May 28–29, 1986 at the Educational Testing Service*. Princeton, NJ: Lawrence Erlbaum Associates.

Byrne, B. M. (2001). *Structural modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Carnap, R. (1956). The methodological character of theoretical concepts. In H. Feigl, & M. Scriven (Eds.), *The foundations of science and the concepts of psychology and psychoanalysis*, (pp. 38–76). Minneapolis, MN: University of Minnesota Press.

Cattell, R. B. (1950). *Personality: A systematic, theoretical, and factual study*. McGraw-Hill: New York.

Cattell, R. B., Eber, H.W., & Tatsuoka, M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire*. Champaign (IL): Institute for Personality and Ability Testing.

Cronbach, L. J. (1971). Test validation. Thorndike, R. L. (Ed.), *Educational measurement*, (pp. 443–507). American Council on Education.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Dowe, P. (2000). *Physical causation*. New York: Cambridge University Press.

Eysenck, H. J. (1953). *The structure of human personality*. New York: Wiley.

Horst, P. (1959). *Differential prediction of academic success*. Seattle: University of Washington.

House, E. R. (1977). *The logic of evaluation argument*. Los Angeles: Center for the Study of Evaluation, University of California (Monograph Series #7).

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement*. Praeger Publishers.

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity*. Information Age Publishing.

Kane, M. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.

Kane, M. (2013b). Validation as a pragmatic scientific activity. *Journal of Educational Measurement*, 50 (1), 115–122.

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Co.

Lissitz, R. W. (Ed.) (2009). *The concept of validity: Revisions, new directions and applications*. Information Age Publishing.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437–448.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences*. Springer Series in Measurement Science and Technology. Springer. The second edition (2023), the e-edition, is open-access.

Mellenbergh, G. J. (2011). *A Conceptual Introduction to Psychometrics: Development, Analysis and Application of Psychological and Educational Tests*. Netherlands: Eleven International.

Messick, S. (1989) Validity. In R. Linn (Ed.), *Educational measurement, 3rd edition*. Washington, DC: American Council on Education.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.

Mislevy, R.J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, New Directions and Applications*. Information Age Publishing.

Mislevy, R., Almond, R., & Lukas, J. (2003a). A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1), i–29.

Mislevy, R., Steinberg, L., & Almond, R. (2003b). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Montgomery, D. B., & Morrison, D. G. (1973). A note on adjusting R^2. *Journal of Finance*, 28, 1009–1013.

Newton, P. E. (2013). Two kinds of argument? *Journal of Educational Measurement*, 50(1), 105–109.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearson, K. (1901). On lines and planes of closest fit to systems of a point in space. *Philosophical Magazine*, 6, 559–572.

Ryle, G. (1949). *The concept of mind*. London: Hutchinson.

Salmon, W. (1998). *Causality and explanation*. New York: Oxford University Press.

Schutz, R. W., & Gessaroli, M. (1991). Use, misuse, and non-use of psychometrics in sport psychology research. In R. Singer, M. Murphey, & L. Tennant (Eds.), *Handbook on research in sport psychology, Chap. 44*. International Society of Sport Psychology.

Schaffer, J. (2000). Causation by disconnection. *Philosophy of Science*, 67, 285–300.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton: Mifflin and Company

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: history repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications*. Information Age Publishing.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, 50(1), 99–104.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement, 3rd edition*, (pp. 263–331). New York: American Council on Education/Macmillan.

Stone, L., & Brosseau, J. D. (1973). Cross-validation of a system for predicting training success of Medex trainees. *Psychological Reports*, 33, 917–918.

Tolman, E. C. (1932). *Purposive behavior in animals and men*. New York: Appleton.

Toulmin, S. E. (1958/2003). *The uses of argument*. Cambridge University Press.

4

Measurement Error and Classical Test Theory

CHAPTER MENU

4.1 Introduction and Overview of "Unfinished Business" from Chapter 3, 154
- 4.1.1 (§40) Domain Theory and Evidence-Centered Design, 154
  - 4.1.1.1 Measurement Instruments and Their Design and Development, 154
  - 4.1.1.2 Domain Theory, 155
  - 4.1.1.3 Assessment in Relation to Measurement, 157
  - 4.1.1.4 Resources for Design and Development, 158
  - 4.1.1.5 How to Access Documentation and Resources for ECD and PADI, 160

4.2 Measurement Functions that Provide Associative Links Between the Empirical World and the Informational World, 160
- 4.2.1 (§41) The Gauss Model for Replications, 164
  - 4.2.1.1 Replications and Random Errors, 164
- 4.2.2 (§42) On Direct, Derived, and Indirect Measurement Across the Sciences, 164
  - 4.2.2.1 What Is Direct Measurement and How Does It Make Indirect Measurement Possible?, 165
  - 4.2.2.2 Concatenation vs Instrumentation's Indirect Results, 165
  - 4.2.2.3 Derived Measurement, 166
- 4.2.3 (§43) The *Specific True Score* τ of a Separate Test/Indicator, 167
  - 4.2.3.1 Test–Retest Method, 168
  - 4.2.3.2 Observed Score, True Score, and Error, 168
- 4.2.4 (§44) The Test as a Measuring Device Compared to the *Precision* of Measurement in Physics and Technology, 169
  - 4.2.4.1 Imprecision and Relative Error, 169
  - 4.2.4.2 European Governmental Standards and Guidelines for Defining Precision, 171

4.3 Reliability and Classical Test Theory, 171
- 4.3.1 (§45) Reliability vs. Precision, 172
- 4.3.2 (§46) Reliability of a Test, Efficiency of Measurement, and the Information Gained, 173
  - 4.3.2.1 Fisher's Information and Efficiency, 174
  - 4.3.2.2 Shannon's Entropy, 174
- 4.3.3 (§47) How to Recover the Unknown True τ? Randomness as the Fundamental Principle in Classical Test Theory, 175
  - 4.3.3.1 Parallel Forms, 176
  - 4.3.3.2 An Application of CT Theory, 176
  - 4.3.3.3 The Expectation Principle, 177
  - 4.3.3.4 The Randomness Principle, 178
- 4.3.4 (§48) Reliability of τ in Terms of the Platonic Conception of CTT, 179
  - 4.3.4.1 Four Axioms of Classical Test Theory, 179
  - 4.3.4.2 Using Classical Test Theory, 180
  - 4.3.4.3 Some Practical Applications of Classical Test Theory, 181
  - 4.3.4.4 Problems in Applying CT Theory to the Reliability of Derived Units, 182

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

4.4 Summary, 182
4.5 Study Questions, 185
References, 189

4.1 Introduction and Overview of "Unfinished Business" from Chapter 3

This fourth chapter is centered on measurement error and classical test theory (CTT). It also continues some themes carried over from Chapter 3. CTT is also called "true-score theory," which assumes that each person has a true score as if there were no errors in measurement. There are always errors in measurement but random errors have an expected value of 0 over an assumed infinite number of replications of the test (τ) that gives the score x. Thus, $x = \tau + 0$ is the expected value. We only have the raw score x, not its expected value, and the error for that person is unknown. CTT gives methods for dealing with these facts. Thus, it is also known as true-score theory. The psychometric techniques for estimating various forms of error and reliability have been greatly elaborated, as you will see in the rest of this chapter (and especially when you go through the example psychometric analyses in Computational Guide #1 on CTT). "Modern" Test Theory uses item response theory (IRT), which is discussed in Chapter 5 (and you will practice implementing this method in Computational Guide #2 on IRT). Today's software packages are much more powerful and general than those mentioned by the senior author, Petr Blahuš, and are available to students and other readers of this book.

There are two main themes of continuing emphasis from Chapter 3 that will be discussed here. The first is about ***measurement instruments*** **and their design and development**. The other is about the ***measurement functions*** **that provide associative links** between the empirical world and the information world. We introduce and use concepts and terms from appropriate and recent theories and models to explain that linkage. This topic is explained in Section 4.2, the introduction to Sections §41–§44.

4.1.1 (§40) Domain Theory and Evidence-Centered Design

4.1.1.1 Measurement Instruments and Their Design and Development

Section 3.1 in Chapter 3 led to the realization that attention to the measurement instrument has been banished from the mainstream definition of validity. This leaves out several decisively important issues associated with what we will group under three topics: (i) *Instrument Design*, testing, and improvement; (ii) *Implementation Design*[1], testing, and improvement; and (iii) *Scaling* of the instrument to an understandable public scale for comparability and ease of interpretation. Despite having defined validity so narrowly that it leaves out these essentials (which the professional standards do not omit), validity theorists like Kane, Messick, and Sireci do advise how to develop evidence for validation arguments during instrument development. We have selected and quoted Kane in the next chapter to speak to it in connection with content aspects of validation. We also cite below

1 Once the instrument is ready for its first pilot administration, and every administration thereafter, a detailed set of procedures for test administration must be ready. Like the instrument, the procedures are expected to evolve during pilot and field tests, and thereafter as needed. Test administration procedures standardize each administration to make it equally fair every time and secure it against cheating and data compromise with continuous monitoring and attention to design.

works on instrument and implementation development generally, then end by featuring the seminal work of Mislevy and others on evidence-centered design (ECD) and model-based reasoning.

Design and development are important to the senior author. In other places in his original manuscript, he mentions the importance of design more directly:

> §71 "*Sometimes the lack of a deeper elaborated theory and a well-founded hypothesis causes the concepts modeled by latent variables to be unclear from the very beginning.*"
>
> §72 The well-known bon mot "*You can rarely put right by statistics what you have messed up by design.*" Also, "*... the design of data analysis has to be an inseparable part of the preparatory stage of a research project.*"

He does not mention structured instrument or implementation design and development methodologies, but he created table 3.3 to present a pattern of progressive development and increasing scientific rigor in instrument development. His table has a general interpretation: as constructs and their measurement are investigated (over many years by many investigators), a family of instruments keeps getting better. The last two levels in table 3.3 could also be interpreted as a design and development progression of a set of constructs and instruments designed and developed by a single investigator or team. In this view of table 3.3, only the more advanced levels would be needed. Instead of the first two steps, literature reviews of past activities in the domain would suffice. Such a project might take years, but not necessarily decades. Promising constructs can be turned into excellent instruments by disciplined and principled (theory-based) development work. These projects involve design, tryout, and revisions using best-practice methods and the best psychometric tools available at the time. The psychometric-methods product is increasingly well-focused and creates excellent instruments that have been oft-revised. Developers should be proud of revisions because they indicate a strong connection with the empirical world (a revision is an indication that improvements were identified following real-world experimentation).

As science progresses and more is understood about progressive attainment over properties of some domains, the institutional use of tests and educational measurement instruments, not just psychometrics (as forecast in Petr's table 3.3), will evolve. Educational organizations will learn how to reorganize to use them optimally, creating different designed implementation plans and ways to organize in place of conventional uses of testing. An example of this is given in Chapter 8, the online, competency-based Western Governors University (WGU). WGU is obsessively student-centered and organizes its uses of measurement and assessment of competencies around what helps individual students progress. They permit multiple attempts at progress-oriented tests, providing feedback and guidance after each testing session. They have separated from the single-teacher role with new mentoring and teaching functions and centralized assessment development, scoring, and feedback. They especially need the evolving science providing explainable theories of progressive attainment. These both provide validation evidence for the claims of scientific definition and comparability; but crucially, they provide guidance for what feedback learners need after each repeatable administration of domain competency exams.

The next step in this account of the centrality of design is to put it in context with something Messick was keen on but was not able to complete during his lifetime—domain theory.

4.1.1.2 Domain Theory

In his brief mention of domain theory, Samuel Messick revealed his growing intuition about where he thought the study of content and construct validity was going. In Messick's (1995) American Psychologist Article, he declared the importance of something he called *domain theory*.

> The boundaries and structure of the construct domain can be addressed by means of job analysis, task analysis, curriculum analysis, and especially domain theory. In other words, *scientific inquiry into the nature of the domain processes and the ways in which they combine to produce effects or outcomes. A major goal of domain theory is to understand the construct-relevant sources of task difficulty*, which then serves as a guide to the rational development and scoring of performance tasks and other assessment formats. At whatever stage of its development, then, domain theory is a primary basis for specifying the boundaries and structure of the construct to be assessed (p. 745, italics added).

See also Messick (1998) for more advice about specifying boundaries and structure.

The importance of this concept is in the implications of the words *Domain-Specific* and *Theory*. Educational and psychological properties that can be measured are initially domain-specific. Once instruments are developed, their generalizability to other domains is a matter of external validation evidence. To acknowledge that the phenomena are domain-specific is to recognize that there is no wide generality in human constructs as in physical measurement scales like length and mass or temperature that apply to objects everywhere and of any size. One can perceive historically an unspoken wistful hope that some properties in the human sciences might be found which are also very general. But even seemingly stable personality traits can change with time, strong experiences may discourage them, and conscious effort can form new habits.

Skills and abilities of people, the stuff of educational measurement, require extensive domain analyses and models for cognitive and/or physical processes that represent what people can perform or can learn. Starting with the idea of four types of construct validity does not help find the domain-specific dimensions or the level of attainment. It requires specific language and diagrams—meanings that person domain models of these skills and abilities provide. Next, how do we design empirical tasks that capture what the models say people can do? The task models need the student progression models first, in order to develop the empirical tasks that probe the distinct differences between the distinguishable levels of progressive attainment. Thus, there is more complexity. It is not just domain specificity; it is also definition of levels along the pathway to excellence of what may prove to be a multidimensional set of scales.

Where the emphasis in the last paragraph and elsewhere is on educational measurement, it is because educational, professional, and occupational testing have created the largest testing companies. The companies will have pressure to refocus if market conditions point the way. This is happening with the rapid growth of WGU, built around learner-centered assessment which demands repeated assessment followed by guidance and coaching. This refocus is not just applicable to educational measurement. The use of new kinds of tests in clinical settings in psychology; sociology; audiology; physical, occupational, and speech therapy; for example; also have scientific models of progression into (and out of) conditions that therapists treat and processes that scientists study. They also can benefit from forms of testing focused on progress, in place of testing used as a static snapshot, for admission to (or graduation from) therapy. The scientists studying these human domains can also benefit from the methods and concepts in this book.

The superdomains of extensive and intensive physical properties (see Chapter 8) do not have the long developmental and practice periods that humans have. It is learning, with persistent effort and belief, that enables individuals to attain the highest levels of excellence. There are physical analogies (such as the time-consuming geological processes that produce diamonds, and other gemstones, after long years of pressure, temperature, and favorable chemical conditions) but no gemstone rose to excellence due to its own volition. Measurement in the human sciences has more varied domains and more levels of attainment than can ever be comprehensively addressed.

It is wise to select the domains with most value to the most people, as well as the high-value ones for coaching to exceptional excellence.

Messick's intuitions about the importance of domain theory were wise. He was speaking about domain-specific theory being the primary basis for specifying the boundaries and structure of the construct to be assessed. He was incomplete, however, in limiting his emphasis to finding *construct-relevant sources of task difficulty*. The previous paragraphs are meant to show that it is the person theory or model that precedes and guides the task theory or model. Together with the evidence model, they truly define the boundaries and structure, and how to interpret and use the meanings of scale values from an educational or psychological assessment event.[2]

Other Uses of the Term "Domain Theory" We will briefly touch on how "domain theory" is used in Mathematics and Computer Science, prior to delving into how it is used in Psychometrics. Messick did not mention that the term "domain theory" had been used for decades in mathematical studies of partially ordered sets (posets) commonly called domains. Computer scientists have found mathematical domain theory very useful in areas such as finite automata, which are not governed by determinism. Future scientists who regard human choices and behavior as nondeterministic and bounded within partially ordered sets lacking infinite continuity may find future uses for mathematical domain theory.

Some psychometricians have used the term "behavioral domain theory" (see McDonald, 2003; Marcus & Borsboom, 2013, chapter 5). McDonald (2003) conceived a behavioral domain as an infinite collection of *items* that are homogeneous to a behavioral construct. Infinite domains (at least nonmathematical ones) of human-curated items seem nonconstructible and unreal. Markus and Borsboom critiqued behavioral domain theory in Chapter 5 of their *Frontiers of Validity Theory*. They ask whether another kind of behavioral domain-specific theory, a Causal Measurement Theory, could also exist, and give pros and cons of such a theory and its relationship to validity theory. Presumably, they would not use the term "behavioral domain theory." Nor will we again in this book.

4.1.1.3 Assessment in Relation to Measurement

Assessment is a word from human measurement activities. It includes (but is more than) measurement instrument development and use. By enabling reasoning from the models and the measurement results, assessment is also the means to make correct interpretations of different values on a measurement scale. By having evidence for claims, decisions are not likely to be overturned by legal actions. For high stakes assessments that pass or fail individuals, the evidence for a valid cut-off score (along with other pertinent non-test information) justifies the action of either failing or passing and thus gaining admission to a desired educational program, attainment of a certificate or some other desired outcome. Failure shuts doors, at least until the test taker can get ready, pay again, and take it again. Correct interpretations of assessment instruments that are administered repeatedly, enhance the learning process by enabling a mentor to give level of attainment-appropriate feedback or coaching suggestions. For example, when the assessment indicates that a learner is ready to take the assessment for the last time, the mentor has justification to encourage the learner to muster what it takes to get to the next significant level.

---

2 The use of the terms for Person, Task, and Evidence models is in support of the author's claim below that evidence-centered design has provided a design language. Common terminology must be based on a comprehensive analysis of the whole process of instrument development, including validation evidence for the instrument's claims. ECD has given terms and meaning for design concepts in the fields of psychometrics and instructional design that can serve people in varied roles and professions to converse about the process of design.

In high stakes testing practices, the most uniquely human characteristics have been sidelined: learning capacity, persistence, and belief in themselves. These growth-promoting aspects of educational measurement domains need to be placed at the center, rather than forgotten, in order to better inform the decisions institutions make that may otherwise be valuable to their employed incumbents but not to their students. Similar advice was given at the end of *The Handbook of Test Development, Vol. 1* edited by Steven Downing and Thomas Haladyna (2006). Cynthia Board Schmeiser of ACT, one of the important developers of the ACT college admission tests, wrote the epilogue. After musing that testing alone will not accomplish educational reform, she remarks "*One wonders, however, whether we as test development professionals need to focus more on the design of tests that are directly connected to instruction as a fundamental component of test design so that they not only provide more detailed information about student knowledge, skills, and abilities, but provide their results in term of empirically supported instructional interventions*" (Downing and Haladyna, 2006, p. 760).

4.1.1.4 Resources for Design and Development

The reference above to *The Handbook of Test Development* is a good place to launch this topic. The author quoted above also comments on why such a volume was important. She points out that the intent of the entire handbook is to replace with empirically supported practice "... what has become the 'common law' of test development conveyed through advice and opinion through its history" (in Downing and Haladyna, Eds, Handbook of Testing, 2006, p. 759). Schmeiser's pithy characterization of test development companies' practices as "common law"—a collection of shared experiences and opinions—resonates with Petr Blahuš's warnings. Over his experience and career, he saw many scientifically inadequate "common law" practices and made strong efforts to convey his vision for what would be better.

The continuing need for such ready documentation of best practices in one book led to the publication of Volume II of the *Handbook of Test Development* (Lane, Raymond, & Haladyna, 2016). It is interesting to note that Mislevy and Riconscente (2006) supplied a chapter in Volume 1 and Riconscente, Mislevy, and Corrigan (2016) provided another in Volume 2. Both were titled *Evidence-Centered Design*, and both were early chapters under the heading, **Foundations**.

This placement of ECD as a foundational element in an upstepping of historically *common law* practices is evidence of how new foundations for something much larger can be placed besides older foundational errors as though there were no difference. To correct this, we suggest three ordered levels of progression toward scientific respectability, something in sympathy with what Petr's table 3.3 attempted to do. Relying on the Joint Standards and on quotes from Nobel laureate Herbert Simon's concept that there is a science of design, we list three levels for human sciences assessment practice.

Traditional Trade Practices in Testing Company Development Programs There is a broad range of quality in these programs, but they do not always meet the Joint Standards. Herbert Simon (1984, p. 156) had something to say about this level of design at its worst "But nowhere do we need to retreat to the methods of the cookbook that originally put design into disrepute and drove it from the engineering curriculum."

Upstepping to Higher, more Scientific Standards *The Handbook of Test Development* is built around the Joint Standards, and best practices are shared in numerous other ways. Sessions in conferences of the National Council of Measurement in Education (NCME) and the Association for Test Publishers (ATP) provide forums for practitioners to share their best practices. Organizations like these can also play an important role in Legal Metrology, discussed in Chapter 8. As an example, ATP

provides a pdf of their successful advocacy to protect test publishers and test users against unwise laws that, for example, force test publishers to share test content with parents or state school officials, thus compromising the integrity of the test. See their pdf summarizing their accomplishments since their founding in 1992. https://www.testpublishers.org/assets/ATP%20Was%20there%20202021.pdf.

Science of Design for Human Sciences A third level could now be defined. It is a step above the standards, and seeks to lead to better scientific, applied, and legal metrological practices in Education and Psychology (as discussed in Chapter 8). Toward this goal, we will again use the words of Herbert Simon (1984, p. 132), "The professional schools (he included education) will resume their professional responsibilities just to the degree that they can discover a science of design, a body of intellectually tough, analytic, partly formalizable, partly empirical, teachable doctrine about the design process." For reasons discussed next, an excellent way to elevate more assessment products to this third level is for more test development operations to adopt ECD as their flexible design framework.

We recommend ECD for multiple reasons. Because of its comprehensive foundational models like the four-process model, and the final layer it covers (in its own way) the five aspects we have recommended for a test instrument's life cycle: *design instruments* → *design instructions for use* (how each administration is managed) → *scale the instruments* → *interpret the results* (from an administration) → *use for each purpose* (the score values and their interpretations).

ECD's principled, teachable, layered design process has proved itself adaptable to a wide range of domains and approaches pursued by involved corporations and higher education institutions and centers. The many published applications of ECD in different domains have a wide co-authorship (people quite literate in what can be done with ECD and who have published, or co-published, on ECD in their own domains of expertise). Add to this the reality that many governmental and corporate procurements for assessment products specify ECD as the required methodology. These factors lead to broad familiarity and use, and constant improvements and innovations come when new groups use it, learn what other groups have done, and contribute their own innovations and slants on the knowledge being gained about learning and performance in their own subject-matter domains. It is good for the understanding of specific domains that they now are receiving a lot of attention by a variety of professional domain analysts and domain modelers.

A common design language, and its generally accepted names for shareable components, can lead to a healthy applied metrology industry. There is evidence in the Principled Assessment Design for Inquiry (PADI)[3] technical reports alone, and in many other publications, that using ECD (increasingly facilitated by the online PADI system and other systems that fit the ECD layers concept) enables groups and individual researchers to develop and test instruments and applications, and share their design patterns and other components. This facilitates Applied Metrology—analogous to the building construction industries that share common terminology and familiar parts like two-by fours, 4 × 8 and other sized panels (with materials ranging from wall-board and plywood to granite and exotic wood panels). For these and other reasons, from this point in history, ECD seems a very promising vehicle for both enhancing scientific metrology in the human sciences, and producing a flourishing applied, industrial, and legal metrology as already exists in the STEM[4] fields.

3 Principled Assessment Design for Inquiry (PADI) is both a library of technical reports and a set of online software tools for developing assessments aimed at teaching inquiry in science education.

4 STEM is an acronym for Science, Technology, Engineering, and Mathematics.

4.1.1.5 How to Access Documentation and Resources for ECD and PADI

The ECD literature is now extensive, but we are also suggesting reports available from the SRI website on the PADI project. A good place to start is an attractively packaged technical report authored by Mislevy and Riconscente (2005) titled *Evidence-Centered Assessment Design: Layers, Structures, and Terminology*. Not only does this report explain very well the terms in this title, but it also provides references to a wealth of other resources on ECD in the final appendices at the end of the report. The aforementioned report is actually PADI Technical Report #9 in a series of reports published by SRI International as principal investigator on the multi-year PADI project. The foundational report, explaining the concept of the PADI online software from its conception, is found in PADI Report #1, and a summary of Frequently Asked Questions about Using the PADI System software is found in PADI Report #12. Of particular interest in this book is the report #10 On Learning Progressions. It explains what learning progressions are and provides a Student Model viewpoint for them. Any ECD example is domain specific, but learning progressions are similar to what was called Domain Theories of Progressive Attainment in the Brigham Young University group mentioned in the next chapter. The design language of ECD provides methods and terminology more likely of wide acceptance and is very welcome to us.

4.2 Measurement Functions that Provide Associative Links Between the Empirical World and the Informational World

At the time his manuscript was written, Petr Blahuš was influenced by the writings of Ellis (1966) and by his Czech colleague Karel Berka (1983) about the idea of "associative measurement." He also mentions Krantz, Luce, Suppes, and Tversky (1971) on associative measurement. However, Kranz and colleagues did not recognize Associative Measurement as a separate category of measurement. Instead, they see Ellis as its proponent and discuss his two chapters on it. They consider it as a special case of conjoint measurement (Krantz et al., 1971, p. 503).

Petr, however, was using it for providing scientific associative links between the observable world in the left side of table 3.3 and the unobservable-but-theory-based explanatory world on the right side. He initiated the need for this starting at the end of Chapter 3, and elaborated it in Chapters 4–6. We do not find that Ellis' original notion of Associative Measurement has attained any kind of following in the 15–18 years since Petr's manuscript was initially written. But his intuitions for what was needed to establish scientific measurement were sound. Table 3.3 is both a model of progressive improvement, from top to bottom, both of constructs and of individual instruments that have been developed to attain high scientific evidence and authority.

As discussed above, the progress in instrument development in those 15–18 years has been notable. We have cited the milestones of the Vols. 1 and 2 of *The Handbook of Test Development*, and as it is linked to the standards, they trace improvements in the standards themselves, which contain guidance for best practices in test development. We have also explained the importance of ECD and its software support tool PADI for the difficult measurement of inquiry skills in learning the thought processes of scientists in school settings.

Of great relevance to this book, the Mari, Wilson, and Maul (2021) approach to defining measurement "across the sciences" contains what Petr and many others of us have been looking for. It provides a well-developed and seminal contribution to measurement across all the sciences, and these foundations have the prospect of being widely accepted and used. To the extent that they are accepted as proposed additions or amendments to the existing concept definitions in

International Vocabulary of Metrology (VIM)[5], they will have an international structure to keep improving them and the applied and legal metrology they stimulate. The lack of common terminology has been a serious vexation for Petr Blahuš. Even as we assume that Messick would have welcomed ECD as better than domain theory, we believe that Petr would be delighted that Mari, Wilson, and Maul (2021) addressed his concerns for associative links from the empirical to the explanatory level and has a hopeful future.

We will next review how Mari, Wilson, and Maul (2021) clarify and accomplish the goal Petr sought to select, among the approaches that he called "causal", "propensity", and "measurement". Petr's original manuscript for this book took a measurement approach, the approach Petr favored, but it also absorbs the causal approach, and does so through the use of an instrument that incorporates strong, new arguments favoring Borsboom's simplified definition of validity. A property must exist, the instrument must have sensors for the property, the sensors must come in contact with the property; this contact must then cause changes in the numerical values delivered by the instrument. Mari et al. have defined and defended strongly that measurement so defined with such an instrument can be seen as direct rather than indirect.

We will use the Mari et al. definition and refer to the stages the instrument goes through as an empirical and informational process (as shown in the black box depiction of direct measurement on page 37 of Mari et al.) This develops the definition and illustrates it as a black box, containing an instrument, that accomplishes three linking functions. This is a specific instance of how the Mari et al. approach realizes the transitions from the left side – OBSERVABLE (to be explained) – to the right side – UNOBSERVABLE (theory provides explaining). The first clarification of Petr's table is that the observations belong in the *Empirical World* while the explanations reside in the *Informational World*. The black box takes properties of objects in the empirical world and outputs values on a publicly understandable scale into the information world. As Petr used the term "explaining," his table invites the more useful term "informational world". There is nothing mysterious or unobservable about the numerical values produced by an instrument. These values result from observable properties of real-world objects interacting with the instrument's sensors, which convert those properties into interpretable values on a publicly meaningful scale—understandable by users within the relevant domain. In addition to assigning individual-specific predicted or estimated scores as the result of a particular latent variable model scoring and calibration activity, we note that many of the important investigations into the distributions and relationships among scientific constructs remain at the latent level without ever attaching or assigning model-derived values or scores to any individual. If the testing market grows ever more user-centered, as we predict, the test takers will deserve as much valuable feedback as the organizations using tests can provide.

In Section 4.1, we have explained the availability of formal design and development processes that instrument developers can use to analyze the domain and develop the models and interpretive information needed to use the output in the "informational world." This is how they "*design on purpose*" valid measurement instruments, starting with properties of objects. We now have the first part of the definition of measurement. It is this: *Measurement is an empirical and informational process designed on purpose, whose input is a property of an object.*

5 The VIM as it stands is insufficient for all sciences because (1) it deals with physical properties in a broad sense, including chemical, biological, and medical. (2) It has a strong focus on "quantities", usually ratio scales. For these two reasons it needs to be broadened to fit the needs of all sciences. (3) Because of the structure of the VIM as a vocabulary, specified by ISO regulations, it has structural requirements for the *intention* of a definition, making it narrowly hierarchical, which sometimes leads to contradictions. This required a special set of notes and conventions to present the *For All Sciences* definitions for consideration in future VIM editions.

Mari et al. (2021) provide a black box diagram on page 37 that explains their definition. Rather than reproducing the black box diagram, we will outline two lists, one for what happened to the inputs on the empirical world side, which moves from left to right, and another that explains what happens to them inside the instrument to produce the numerical values to be used on the information world side. There are three processes inside the instrument, with the first one *Transduction* feeding an output to the middle one *Instrument Scale Application* producing the first output in the *Information World* side and feeding it to the last process designed on purpose—*Calibration*, producing the final value in the information world.

Empirical world (left side of the black box diagram Mari et al., 2021, p. 37).

Transduction process. People using the test instrument(s) bring an object (a person or the person's vector of responses) in contact with a test instrument so that its "sensors" can obtain information which will cause it to produce an internal *instrument indication*. Remember that the existing definitions for physical measurement must be preserved, as well as broadened to other sciences not embraced by the current VIM. Sensors of all kinds and forms are used in physical measurement. Tests or rating scales are the common form of sensors in most human assessments, but a bar that gets cleared or knocked off in high jumping or pole vaulting is also a sensor. So is touching the dive-off wall at the end of a swimming race to trigger the time clock. Online testing apps can present the questions or receive ratings from judges, and apply the relevant transduction process. Educators have for years called it scoring, and probably will continue to do so. The result of this is called the *instrument indication*.

We explained the empirical process of *Transduction* in Chapter 3, illustrated by the thermometer and a test of Reading Comprehension Ability—two quite different mechanisms for transduction. Transduction is domain-specific and depends on the scientific models appropriate to each domain. At level b of instrument development in Petr's table 3.3, it is simply giving a 1 or 0 for each question and adding them up. The result is called the *instrument indication* in metrology language and is the sum of the 1s and 0s—called the raw score x in CTT. Doing calculations like averaging a set of raw scores for a group of people scored in this manner ignores the fact that such raw scores are only ordinal—another reason why this level of work is at level *b*. But the new definitions of measurement will permit this. This is characteristic of Petr's level *b*.

But at level *d* of table 3.3 of a human testing instrument, the transduction function used by the instrument could be much more powerful. What is called "the instrument" contains in its calculation programming the parameters of difficulty for each question, estimated ideally using data from thousands of representative persons. Various IRT models are available, and latent factor models as well. Given these nonlinear parameter estimates, the instrument is able to calculate the person's score values on an interval scale (assuming uncorrelated errors). It can also maintain the standard error needed to report uncertainty in the value reported. Even at this advanced level *d* in Petr's table, the result of transduction is called the *instrument indication*. This intermediate representation goes next into the process that links between the empirical world and the information world. This next process following transduction is called *instrument scale application*. Since the output of this process is a numerical value of use in the information world, it will be discussed next under that heading.

Informational World Functions of the Instrument These functions complete the Mari, Wilson, Maul (2021) definition of measurement ... *and that produces justified information in the form of values of that property*.

The *instrument scale application* process is the bridge between the empirical world and the information world. It touches both sides, so what goes in as an *instrument indication* on the empirical side comes out on the other side as the *indication value*. It is now a part of the *information world*. This transformation from *instrument indication* to *indication value* is all part of the middle subprocess called *instrument scale application*. This term was used in Chapter 3 where it was recognized as the second "associative link" that satisfies what Petr was looking for in table 3.3. There may be no difference in educational measurement between the instrument indication and the indication value, but in physical measurement, the instrument indication may be levels of voltage or the level of mercury in a thermometer that has to be matched to the readings on the scale of the thermometer by a human observer. If so, this additional step of *matching* by a human is a part of the *instrument scale application* and the temperature instrument, a glass thermometer that has no calculation logic or memory in it, could not itself deliver the *instrument value*.

In scoring for IRT, the indication value is indeed on a local scale. Most IRT scales use logit units (based on the logarithmic function used to fit the nonlinear item curves, see Chapter 5). This logit value corresponds to the *indication value*. It is not yet the final public value, which we will explain below.

Do not be confused by this usually unnecessary matching step in a nondigital instrument's local scale. Both the glass thermometer and the usual voltmeter need a human to read the value. These low-cost instruments do not contain memory or digital logic to go any further. In both cases, the human reading is usually the final value needed anyway, with no calibration needed. However, there is usually one more vitally important *calibration* step before the person and other users can be shown their results. Many domains have established a particular public scale for interpretation and use. Guidelines and training, for people (employed in the domain) to interpret and use that information, cannot be developed unless a standardized scale exists, and any local scales different from it must be calibrated, assuring that the final informational value is presented on the public scale. Calibration is a vital part of the quality assurance process in international metrology for units accepted as part of the SI system. All SI measurements must be on the same final public scale.

Thus, we have three processes—transduction, instrument scale application, and calibration, which were defined by Mari et al. (2021) in a language appropriate to the VIM. Transduction is fully in the empirical world, calibration is fully in the informational world, and instrument scale application bridges the two worlds. Together they provide the definition of direct measurement for all sciences. These links will hereafter in this book be called by their own VIM-proposed names rather than by Petr's terminology as associative links. Authoritative and unambiguous terminology is a premium in his writings.

We have tried to remove all references to associative measurement. In addition to the conclusions of the present continuing authors, and the Kranz et al. authors, Mari et al. (2021) devote eight pages (section 7.2) in a masterful attempt to resolve historical ambiguities and circularities in the meanings of direct and indirect measurements. They proposed removing a circularity in a published definition of direct measurement with these words. "A measurement method is direct if the measurement instrument is designed to empirically interact with the properties of the same kind as the measurand." Their reading of Ellis is that he understood this kind of directness but chose instead to propose associative measurement as a form of nonfundamental, and thus indirect, measurement. While disagreeing with him, Mari et al. (2021) (p. 198), state that "Ellis may have been the first to include it (the term "direct measurement") in a structural analysis of measurement."

This completes the discussion of the two themes carried over from Chapter 3, and we proceed next with Petr's presentation of Classical or True Score Test Theory.

4.2.1 (§41) The Gauss Model for Replications

The incorporation of errors in measurement is typically treated by an additive model known as the *Gauss measurement-errors model*:

$$x = \tau + \varepsilon$$
$$Measured = True + Error$$

Note that from the point of view of the researcher, the true score τ is an *unknown target* to be estimated from the known x. However, at the same time, the Gauss equation explicitly expressing x seems to work *as if* τ *were known*. This bidirectional *measurement link* between the *known* and *unknown* is inherently contained in any conception and model of measurement. In the Gauss model, both directions of the link have a direct and very simple form—actually taking on a form of a "stochastic identity," up to the random errors. On the other hand, in some situations, the measurement models can become quite complicated functions, as we will see later.

4.2.1.1 Replications and Random Errors

From an inferential statistical point of view, we can imagine replications as sample realizations of a random variable in random samples drawn from an infinite population with an unknown population mean τ. Broadly speaking, the requirement for the postulates to be true is *randomness of the errors* ε in the replications. Randomness of the errors ε allows us to use laws of probability theory and mathematical statistics to derive practical consequences of the model (in this narrower statistical view, the distribution is normal by the conception itself). This independence is a fundamental property, which is implied by the stochastic independence and distribution identity postulates mentioned above. Then, by the central limit theorem, the mean of the replications $\overline{x}$ is governed by the normal distribution and converges to τ as its unbiased estimator $\overline{x} = \hat{\tau}$, and with the increasing number n of replications $n \longrightarrow \infty$, it holds that the mean approaches the true value $\overline{x} \longrightarrow \tau$.

As a consequence of the unbiasedness of $\overline{x} = \hat{\tau}$, the standard deviation s_x converges to the standard deviation s_e, well known as the *standard error* of measurement s_e. In addition, as $n \to \infty$, it itself also vanishes, namely $s_e \to 0$. A very important by-product of this mathematical theory is that it means that, in terms of the Gauss model, the true τ value becomes *statistically estimable*, though it is *not* "statistically observable"—in the sense of inferential statistics.[6]

In summary, true replications are generally impossible to achieve in practice. Replications can only be approximated through repeated measurements or through simultaneous readings by multiple "sensors" or tests. If the errors are random, approximately replicated measurements can help us obtain an estimate of the true value and an estimate of the error variability.

These assumptions work out well enough to have built the True Score Theory edifice of CTT and later IRT, as we will see.

4.2.2 (§42) On Direct, Derived, and Indirect Measurement Across the Sciences

We rely heavily here on the main links between the empirical world and the informational world as defined and explained by Mari et al. (2021). The three main links are Transduction, Instrument Scale Application, and Calibration. Petr defined educational and psychological measurements as indirect. This word remains in Figure 3.3. In the following three sections (42a, b, and c) we have discussed aspects of the definitions of Direct, Indirect, and Derived in terms of the logic of the new definition of measurement from Mari, Wilson, & Maul (2021). Petr's work on calibration is evidence

6 The reader is referred back to Chapter 1, and especially section §13, for student-tested explanation of these concepts.

of the soundness of his intuitions concerning associative links between empirical and formal representations. However, his initial analysis employed mathematical notation that is unused elsewhere in this book (or in familiar psychometric texts), so we have not used it, but have incorporated its intent with reference to the discussion of temperature and calibration from Mari et al. (2021) as discussed above and in Chapter 3. We will abbreviate references as Mari et al. in the next section.

4.2.2.1 What Is Direct Measurement and How Does It Make Indirect Measurement Possible?

In Mari et al., measurement is carefully defined over many chapters and culminates with the definition of *direct measurement*. This careful development and definition void the usual assumption that if a thing is unobservable, the measurement must be indirect. An instrument is a part of the definition, and there are numerous examples of instruments in STEM domains that have sensors that make contact with the property being measured. So long as the instrument makes contact with the observables, and there is only a single dimension, the measurement can be considered direct. By putting the transduction, instrument scale application, and calibration processes inside of an instrument, it meets their thorough and complete definition of direct measurement.

Mari et al. also define indirect measurement, and it includes what writers of Petr's time were calling "derived measurement." The VIM does not define *derived measurement* as a kind of measurement, but talks about derived quantities and derived units. As in the simple example of density, the measured values of two units are needed, where density = mass/volume. Volume is itself a derived unit composed of height, width, and depth. In consequence, in density, there could be four direct measures including a multiplication of the three sides of the volume box. If the measurement uses the volume of displaced water, there would be only two values. When direct measurement methods are used to provide the values of mass and volume, the calculation of division occurs, deriving the value of the density unit. In either case, mass divided by volume produces a derived unit, and because it is made up of two or more direct measurements, and a calculation, it meets the definition of indirect measurement. It is all done by mathematical equations and the measurement is defined as indirect and the unit as derived. It also takes the two extensive physical properties (mass and volume) and delivers an intensive property (density) that does not depend on the size directly. Extensive and intensive properties are explained in more detail in Chapter 8.

In Chapter 3, we discussed and rejected the implication that instruments are not a part of the validity concept. We disputed that the definition of validity could leave out measurement and only apply to interpretations of scores after an administration and their use for some purpose. It is science advancing to put instruments front and center. The advances in technology make it possible for instruments (some are called apps today) to serve as quite complex sensors or probes in the empirical world. Such instruments also have built in access to memory, calculations, time, and GPS locations. Direct measurement of single dimensions so defined is reasonable and demonstrable. So is indirect measurement with complex instruments measuring several SI dimensions at one moment in time. Indirect measurement as something vague and unseen goes away—less of an embarrassment to scientists in the human domains as it becomes both more empirical and more formal, an aspiration Petr sought, described in Chapter 3.

4.2.2.2 Concatenation vs Instrumentation's Indirect Results

"Physical manipulation of rods, for measuring length, was regarded by many, Petr included, as an example of direct measurement. Actually, using rods and concatenation has proven useful for philosophers of measurement to think about, but is not useful as a practical tool for performing valuable work in the world. Instruments that use more than one direct measurement, such as the second and the meter, and produce indirect results, are far more practical. One example is land

survey measurement using a total station theodolite instrument with GPS. Since the space age, such an instrument can give the angles from satellites to different reference points on the ground. This enables the surveyor at a moment in time to place the stakes accurately at the corners and lines of a property. The trigonometry that calculates indirectly the distance between two points on the ground in relation to several satellites within range is ever so much faster, more flexible, and more accurate than the direct methods they replace. It is an indirect, derived measurement involving multiple length measures and time measures for the exact location of the satellites in range.[7]

Because of these realities, *an instrument with valid calculations built in uses indirect measurement. The value produced may be a derived unit. If there is just one dimension being measured, such a measurement is direct*. Measurement across the sciences presupposes this. It also follows for IRT and latent factor calculations for scaling, as well as calibration calculations. Mari et al. deal with uncertainty and what is needed in direct or indirect measurement to enable the instrument to derive final values with known errors of estimate. It is wise and appropriate that such instruments in the human sciences need to be continually evaluated and improved.

4.2.2.3 Derived Measurement

This term used by Petr, quoting various sources, does not appear in the VIM3. There is no recognized special kind of measurement called *derived measurement* in international metrology. There are, however, definitions of *derived quantity*, *derived unit*, and *coherent derived unit*. A derived quantity is defined in terms of its base quantities. A derived unit is the measurement unit for a derived quantity, and a coherent derived unit is defined in a given system of quantities as a product of powers of base units with no other proportionality factor than 1.0. As for the definition of the calculation function, the term "*unit equation*" is found in definition 1.23, p. 34 of VIM. It refers to the mathematical relationship relating to base units, coherent derived units, or other measurement units.

Defining derived as a type of measurement illustrates the lack of citation from international metrology sources found in much of the literature on psychology and educational measurement.

Derived quantities, units, and coherent derived units have a long history. Myriad examples of derived units exist in the coherent SI system of units. Each unit has a dimensional interpretation, and where there are SI units, and even without them, derived units represent a very common activity. Industrial or applied metrology provides the calibrated instruments.

Table 4.1 was table 5 in the BIPD official brochure from 2019. It includes examples of coherent derived units from base units and other derived units. Note that the several forms of density are intensive properties in that they do not affect the extent of the construct being measured. Dividing by area or volume or kg produces intensive quantities.

These 13 coherent derived units arise from extensive or intensive quantities. These are called coherent derived units because these are all SI units derived from products of powers of base units with no proportionality factor (i.e., they follow the definition of a coherent derived unit). Unit sizes are fixed internationally, and a scale value is a multiple of that unit. One does not need to make scale adjustments to enable them to fit together, as in going from Celsius, a non-SI unit, to Kelvin, a SI unit. Coherence is a great benefit, and quite amazing that so many important units in physics, chemistry, engineering, and some of medicine rely on common units in a coherent system of units.

In the most recent SI brochure, a list of 27 derived units is displayed on p. 139. This note is given alongside the list. *A previous resolution "... specified other units which should be added to the list. In principle, this list of derived units is without limit."*

---

7 A homey example of a tool that sometimes involved concatenation of different lengths, and no home builder could be without, is the carpenters retractable roll of calibrated metal tape. The carpenter provides the transduction function by putting the tape in contact with the object and notes the numbers, and then cuts to that length.

Table 4.1 Examples of coherent derived units in the SI expressed in terms of base units, a photocopy of table 5 from p. 135 of *The International System of Units (SI)*. This SI Brochure is distributed under the terms of the Creative Commons Attribution 3.0 IGO.

Derived quantity	Typical symbol of quantity	Derived unit expressed in terms of base units
Area	A	m^2
Volume	V	m^3
Speed, velocity	v	$\mathrm{m\,s}^{-1}$
Acceleration	a	$\mathrm{m\,s}^{-2}$
Wavenumber	σ	m^{-1}
Density, mass density	ρ	$\mathrm{kg\,m}^{-3}$
Surface density	ρ_A	$\mathrm{kg\,m}^{-2}$
Specific volume	ν	$\mathrm{m}^3\,\mathrm{kg}^{-1}$
Current density	j	$\mathrm{A\,m}^{-2}$
Magnetic field strength	H	$\mathrm{A\,m}^{-1}$
Amount of substance concentration	c	$\mathrm{mol\,m}^{-3}$
Mass concentration	ρ, γ	$\mathrm{kg\,m}^{-3}$
Luminance	L_v	$\mathrm{cd\,m}^{-2}$

4.2.3 (§43) The *Specific True Score* τ of a Separate Test/Indicator

In their well-known monograph, Lord and Novick (1968) explain the specific true score by *distinguishing it from* "The *generic* true score defined in terms of a whole family of *different* test forms ..." (ib., p. 174, italics added), say different items measuring the same ability—the authors present an example of "spelling ability" (ib., p. 174). As another example, consider a specific true score and its specific reliability in kinesiology: we can imagine an isolated test consisting of a complex sensorimotor task, whose score is expressed by time in seconds. Then we are interested solely in a possible standard error in the metric of seconds—but we are not interested in the problem of whether the test can reliably estimate the general level of sensorimotor coordination ability.

This narrow look at specific reliability can be further illustrated by an idea about synchronic approximation to replications with an example of four thermometers. Let us assume that we use thermometers of the same kind, i.e. based on the *same principle*. Say, we use four electric-resistance thermometers of the same make, produced by the same factory, and we use them simultaneously. Now, we could have a suspicion that the thermometers based on this specific kind of physical principle and corresponding specific operations measure their own *specific* "*kind of a temperature*," quite different from the other thermometers based on different principles. Such a view would be in accordance with extreme operationism, in which a concept is defined by its specific measuring operations. Therefore, we have the expression "*a* temperature"—we would have as many temperatures as there are different methods for their measurement. The specific features of different thermometers do not mean that their values are shifted necessarily by a systematic bias but, for instance, it may be that their variance is specifically distorted, or a certain kind of thermometer may have a tendency to overestimate high values while it may be quite acceptable for average and low values, etc. Of course, if we knew the specific properties of each principle of the thermometers, we might intentionally use them in different practical situations.

4.2.3.1 Test–Retest Method

With regard to diachronic approximation to replications (or replications over the course of time), we hardly ever find a behavioral diagnostic method that might be meaningfully repeated more than two or three times.[8] However, there are some exceptions. Consider the measurement of simple reaction time. After "warming up" to the task and environment, a participant may be tested repeatedly until the subject becomes fatigued. On the other hand, the V-fluency test (producing names) perhaps might be repeated the next day. Due to the rather low number of meaningfully repeatable testing sessions, the traditional version of CT theory, usually presented in a somewhat simplistic way in older textbooks, is to repeat the test just *twice*, $n = 2$, but in a large set of N *different* individuals.

We may compare the situation to the measurement of the length of N different gardens twice with the same yardstick, under the same meteorological conditions. An analog to establishing precision of a certain kind of thermometer would be to measure temperature twice ($n = 2$) but in N different rooms with different values of temperature, when the temperature in each of them was kept stable in time over the $n = 2$ repeated measurements. These two examples illustrate a simulation of "over-time replications," as a situation in which specific reliability is approximated by the so-called "*test-retest*" *method*. This represents a substantial conceptual switch from the Gauss model of precision based on replications to the formally similar CTT model of specific reliability, namely:

1) In the Gauss model, the unknown precise value is *a constant value*. Hence, if the symbol τ had been possibly thought of in an unspecified sense as a variable, then the Gauss model in Section §45 should have been written in a better way for one *constant* τ^* and *two variables* x, ε: $x|\tau^* = \tau^* + \varepsilon$. The values of the variable x are obtained by a measurement procedure over the random sample of n replications with the given fixed value τ^* (however, it is unknown). Thus, x is a variable conditionally observed under the given condition of the fixed value τ^*, while ε is the corresponding difference.
2) In the CTT model, the unknown precise τ is a *variable* taking its values τ_i over N subjects $i = 1, ..., N$ (not necessarily a random sample). By analogy to the Gaussian conception, we could say that we have just $n = 2$ replications as variables x_1, x_2 taking their values x_{1_i}, x_{2_i} over those N subjects.

4.2.3.2 Observed Score, True Score, and Error

By the tradition of CTT, a somewhat specific terminology is in use for the *three variables*: x is called the *observed score*, τ is the *true score*, and ε is *error*. It is necessary to underline that of the two unknown variables τ and ε, τ is a *specific true score* in the sense of the above-mentioned idiosyncratic identification of the given test and its distinction from the other tests (compare the distinction of mercury vs. bimetallic thermometers, say) and ε is the test error due to the *specific unreliability* of just this given test. It does not have anything to do with an error of measurement of some possible theoretical concept about which the test might also yield some information (as its prospective indicator). This error ε should be *distinguished from the d* mentioned in the preceding Section §46,

8 As discussed in Chapter 8, repeated attempts followed by coaching are optimal for learning in sports and can also be used in academic subjects where the instruments are designed to assess progress up a scale. This is a growing area in assessment. Petr is here discussing finding a single correct measurement of relatively fixed properties, not changing ones as in human progress.

where *d* represented the amount of a "misrepresentation" of the true *generic* latent variable *f*, the variable modeling a theoretical concept. There, this *f* was generic, jointly measured by different indicators, and its values were distorted by *d* in a separate indicator *x* due to the different specific character of each of the individual indicators (although all of them partly contributed to the measurement of the same *f*).

4.2.4 (§44) The Test as a Measuring Device Compared to the *Precision* of Measurement in Physics and Technology

We have already encountered the notion of "precision" several times; now let us present its meaning in more detail. Starting from section §28, we have occasionally used several notions that are related to the idea of "precision," namely:

- the standard error of measurement s_ε (distinguishing this s_ε from the next s_e),
- the standard error of regression estimate $s_e = s_{y|x}$ (sections §28 and §31),
- the relative decrement of the $s_{y|x}$ error by improving a regression equation (sections §31 and §34),
- the coefficient of determination $R^2_{y|x}$ as a proportion of "explained" variance vs. the coefficient of nondetermination as the proportion of "unexplained" variance $s_e = s_{y|x}$, i.e. variance of the errors of the regression estimates (section §34),
- *Fisher's information* $I = 1/s_e^2$, a concept reciprocal to the variance of the errors.

Standardization requirements for the various kinds of measurements distinguish different types of measurement errors. One of the most important distinctions is between systematic error (that is, bias) and random errors (e.g. Federer, 1973, p. 23). Random errors can be expressed in terms of their variance around the mean of the instances of measurement—this represents *precision*. However, high precision itself need not mean a general *accuracy*, as the mean may be systematically shifted below or above the unknown true value, which is the desired target. Hence, *accuracy* is influenced by *bias* as well as by *imprecision*. The main aspect of precision deals with the variance or range of the replicated measurements $x|\tau^*$ around a given fixed calibration value τ^*—as in the Gauss model. Bias in measurement can sometimes be eliminated by calibration of the measuring device, namely by shifting the zero point of its scale to an appropriate level. This is common in certain educational applications and psychometrics, especially if the researchers or clinicians are just interested in the relative position of a subject's score in the range of a sample or population reference category. A typical application is using standardized *z*-scores where the zero point is shifted to the mean and attention is paid to the deviation scores, i.e. centered scores. In such cases, accuracy is very often more or less directly identified with, and/or substituted by precision.

4.2.4.1 Imprecision and Relative Error

In technology, *precision* is usually inversely expressed as the so-called *relative error*—a measure of *imprecision*. One way to express it is in terms of the variance or standard deviation s_ε of the replicated measurements, as was already introduced above for the Gauss model. The problem is that the true value τ^* is unknown. One way to look at the problem is to estimate with another measuring device, whose precision is much higher, and which is used as a "gold standard." This method is also often used in biomedical laboratories. Of course, the "gold standard" values may also be contaminated by some nonzero errors, though hopefully much smaller: Imagine calibration of a wooden yardstick with respect to the platinum standard known as the *International Prototype Metre*, stored in *Bureau International des Poids et Mesures*, at Sèvres,

France. It used to be recognized as the official meter until 1960 when it was substituted by a more precise definition using the wavelength of a certain kind of radiation measured by an interferometer.[9]

The current definition is based on laser measurement where 1 m is defined indirectly in terms of time, i.e. as the distance the light emitted by the laser travels in a certain fragment of a second. The calibration process of a rather rough field measurement method could be outlined as follows. When the gold standard value is obtained as the best possible approximation of τ^* with its high precision, then a long series of n replicated measurements x is carried out. The largest absolute deviation of x from the gold standard is found and considered to be an auxiliary approximation $\tilde{\varepsilon}_{max}$ of *the maximum absolute error*. It is then related to the range $Ra(x)$ of the scale of the measuring device. Thus, an approximate estimate of the *relative maximum error*, called the "*imprecision*" of the device, can be expressed as $Imprec(x)$. It is usually referred to in percentages by $Imprec \times 100\%$. The measuring devices in technology are classified into different categories of precision. Just for illustration, for ordinary measurement in electrotechnics, the lowest precision category of the devices with relative error from 0.03 to 0.05, from 3% to 5%, is acceptable for very rough field measurements.

Currently, the above-presented approach might be considered as an illustration of an out-of-date approach for approximating the precision of a crude field device. In principle, there are at least two approaches for approximating the fixed value τ^*, and consequently approximating (or estimating) the standard deviation s_ε of the errors ε, and their absolute maximum error $\varepsilon_{\max}$:

1) As illustrated above, the τ^* can be approximated by a gold standard value and then numerical deviations of the replications $x|\tau^*$ around it stand for the errors ε, and the estimate of s_ε may be computed (if there is no bias). In this case, τ^* becomes "almost known" if the precision of the gold standard is dramatically higher, say on several orders of magnitude.
2) The τ^* is fixed as an unknown constant, exactly as desired by the formal definition, and this is approximated by certain real-world arrangements. We already had the example of the measurement of a garden (whose length is constant in the sense of practical needs), or the case of an iron rod whose length may be considered constant under stable temperature, etc.

The current state of affairs corresponds to approach (2). Nevertheless, during the last couple of decades, the general approach and terminology of technological measurement have been influenced by governmental offices responsible for standardization, calibration, and comparability. In Europe, special treatment is devoted to the accreditation of different laboratories in the European Union, and this official process started using new terminology, especially by bringing in the new term *uncertainty of measurement*. From the point of view of wider relationships to other scientific branches, this term may cause some problems, however. It might easily be confused with "uncertainty" in the sense of information theory, where it is used to denote Shannon's entropy, which also is related to measurement and its precision—see section §51. Nevertheless, it is necessary to accommodate this situation but clearly distinguish between the two meanings of this homonym.

9 In 1983 the meter's scientific definition was changed again, and reworded in 2019. It is now measured using the distance travelled in one second at the speed of light in a vacuum (see Chapter 8). Petr here gives this partial example of tracing the calibration of an instrument (or the machine stamping out such instruments industrially) to an SI unit.

4.2.4.2 European Governmental Standards and Guidelines for Defining Precision

The process known as the European Accreditation of Laboratories (EAL) and the West European Calibration Cooperation (WECC) in different countries[10] states that an officially recognized measurement laboratory requires that uncertainty be established after "corrections and correction factors for systematic effects" have been done (EAL, 1997, p. 6). It means that a bias is excluded from their concept that represents precision but not accuracy. This *imprecision due to random errors* is then called "Type A uncertainty" of measurement (EAL, 1997, p. 4).

The instances of a measurement procedure yield separate measured values $x|\tau^*$, $j = 1, 2, ..., n$. Although the term "*replications*" is *not* used for the instances of measurement in the materials of EAL, it is assumed that (under moderate statistical postulates) the mean of these values converges to the unknown given value τ^*, and that their variance $s^2_{x|\tau^*}$ may be used as an estimate of the *variance of errors* $s^2_{\varepsilon|\tau^*}$. It means that the EAL assumptions are obviously the same as those two postulates (a, b) used for the stochastic definition of replications (in Section §45). It becomes clear from EAL (1997, p. 13) that:

> 5.2. The assumption of a normal distribution ... derived from well-behaved probability distributions of independent quantities, e.g. normal distributions or rectangular distributions, contribute to ... the conditions of the Central Limit Theorem are met and it can be assumed to a high degree of approximation that the distribution of the output quantity is normal.

Thus, the corresponding standard deviation s_ε can be recognized as the *standard error of measurement* in an older terminology. However, it is now called by EAL (1997) the *standard uncertainty* of measurement, u—but $u \equiv s_\varepsilon$. Further, according to EAL (1997, p. 13), "It has been decided that laboratories accredited by members of the EAL shall [also] state" the so-called *expanded uncertainty* $U = ku$, with a "coverage factor" $k = 2$, which, they say, expands the "coverage probability up to 95%," in their terms. Actually, this "coverage" seems not to be anything other than the interval with confidence $(1 - \alpha) = 0.95$ where a more precise value of k should be $k = 1.96$. In addition, it is also not quite clear why some other values of the "coverage factor" were not recommended. Finally, it follows from the whole EAL conception that their "Type A *relative* standard uncertainty" is the same as the old-fashioned *relative imprecision*, $Imprec(x) \equiv s_\varepsilon/s_x$, possibly also expressed as a percent if multiplied by 100%. Another specialty of EAL (1997, p. 4) is that the precise quantity τ^* of the target variable is called the *measurand*. It is interesting to note the similarity between this EAL terminology and the terms of test theory: The "measurand" of EAL and the "criterion/predictand." (One may speculate whether some terminological and other specialties could be due to the different kinds of experts than statisticians—it seems likely that mostly physicists and engineers were involved in the EAL calibration and standardization process.)

4.3 Reliability and Classical Test Theory

Section §45 describes the relationship between imprecision from the perspective of metrology and unreliability from the perspective of CTT. Section §46 relates these concepts to Fisher's ideas of relative information and efficiency of measurement and discusses the use of Shannon's entropy or uncertainty measures for certain applications. Section §47 emphasizes the importance of

10 These are regional organizations in international metrology discussed in Chapter 8.

probabilistic randomness and the independence of errors as the basis for using parallel test forms as the theoretical and mathematical foundation in the development of CTT. Together, the principles of mathematical expectation and randomness provide a grounding for deriving the four fundamental axioms and the fundamental theorem of CTT in Section §48.

4.3.1 (§45) Reliability vs. Precision

We may see that there are some substantial similarities between the Gauss measurement-errors theory as applied to technological measurement and CTT.

However, there are also some conceptual as well as formal differences. The conceptual difference comes primarily from the fundamental distinction between τ as a *constant* in the Gauss model and/or as the true score *variable* over N individuals in CTT. Some confusion can arise due to the identical term "standard error" and symbol s_ε (also called standard error). These two types of standard errors share a name but have important conceptual and mathematical differences. This represents a nice illustration of the terminological problems—in this case, especially due to the fact that most textbooks use the term "*standard error of measurement*" when the standard error was established according to the conception of CT Theory. Besides these differences, there is still a further one, namely that while the EAL technological notion of "type A *relative* standard uncertainty" or the old-fashioned "relative *imprecision*," is expressed as the *relative standard deviation* of errors, i.e. s_ε/s_x, the CT theory prefers using the *relative variance* of the errors, which is now called the *unreliability*: $Unrel(x) \equiv s_\varepsilon^2/s_x^2$. This proportion of error variance may also be assessed by its complement to 1.00, which is well known as the *coefficient of reliability* or just *reliability*: $Rel(x) = 1 - Unrel(x) = 1 - s_\varepsilon^2/s_x^2$, or $Rel(x) = s_\tau^2/s_x^2$, where the last expression is traditionally used as "*the* definition" of reliability in textbooks.

Using the metaphor (i) of "explanation" by regression and correlation (section §34), we may also follow the concept of the *coefficient of determination* R^2 to say that reliability tells us how much of the observed variability was not contaminated by errors. It means that $Rel(x)$ is the proportion of variance "explained" by the quality of diagnostics. As an extreme case, imagine that the diagnostic method was absolutely incapable of distinguishing between individuals. It would mean that the whole variability of x—that would like to be a "diagnosed" distinction between subjects—would be only due to diagnostic errors.

This approach to formulating reliability allows an interesting comparison between the precision of technological measurement and psychological/behavioral diagnostics. Ignoring their different conceptions, if we apply the definition of *Imprec* (x) to compare with *Unrel* (x), it is clear that

$$Imprec(x) = \sqrt{Unrel(x)}, \text{ and } Imprec(x) = \sqrt{1 - Rel(x)}.$$

Thence, a research diagnostic method with an excellent and rarely found reliability of 0.99 would correspond to a technological measurement device with $Imprec(x) = 0.10$, which means that the reliability $Rel(x) = 0.99$ corresponds to imprecision of 10%. Considering "precision" as a complement to imprecision, we may define it by $Prec(x) = 1 - Imprec(x)$, or in percents by $Prec(x) \times 100\%$. Figure 4.1 presents a graph of that relationship between reliability and precision.

A behavioral diagnostic method whose reliability $Rel(x) = 0.75$ was found by a standardization study that would correspond to a technological device with a relative error of 50%. Of course, the standardization requirements imposed on diagnostic methods in behavioral research cannot be compared to the rigorous standards of technological measurements. Quite often, a technological measurement device is just an inclusive part of the whole machinery of a laboratory examination

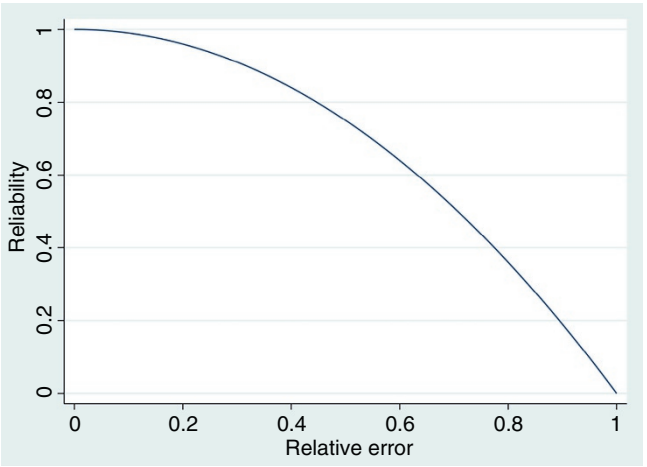


Figure 4.1 Reliability in terms of test theory vs. precision in terms of technological measurement.

and its error is a negligible fragment of the entire complex *diagnostic unreliability*, which contains all of the various possible sources of noise. Depending on the rigorousness of the test instructions described in a reliability manual, it includes all inner and outer factors and conditions. It covers weather, illumination, temperature, instability of the attribute in the tested testee and so on, including the influence of some specific characteristic approach of the administrator/researcher. It is important to distinguish between *two conceptions* of the term "*standard error of measurement*," s_ε—the *narrower*, rather technological conception, and the *wider* one, related to the complexity of the entire diagnostic procedure, which aspect we underline still more by calling it the "*standard error due to unreliability*."

4.3.2 (§46) Reliability of a Test, Efficiency of Measurement, and the Information Gained

The formulation of unreliability as the proportion of the variance of errors of measurement resembles the situation in section §30, which dealt with the proportion of variance of errors of regression estimates. There, Fisher's efficiency estimate was evaluated by $Ef = s_e^2/s_y^2$ while its complement, $1 - Ef$, was interpreted as *relative information* about the dependent variable y yielded by the independent variable. It is quite straightforward to apply the same conception to the problem of reliability by saying that the analogous proportion of variance of measurement errors ε of a test x represents the *efficiency of measurement* $Ef(x) = s_e^2/s_x^2$.

Thus, combining the notions of the preceding section with section §30, we find the simple identity of the Fisherian statistical concepts of efficiency and information with CTT concepts.

"Efficiency of measurement" ... Unreliability: $Ef(x) = Unrel(x)$,
"Relative information" ... Reliability: $1 - Ef(x) = Rel(x)$.

Therefore, also in connection with the preceding sections on precision, we could ask three equivalent questions that differ only in the terminology used:

How precise is the test as a measurement device?
How reliable is the test as a diagnostic method?
How much information does it offer about the unknown values of τ?

For instance, the reliability coefficient $Rel(x) = 0.99$ means that 99% of the observed variance is "true," and 1% is due to errors. Therefore, the precision (relative maximum error) is $\sqrt{1-(0.99)} \times 100\% = 10\%$, and our auxiliary index of "Relative information" is 0.90.

4.3.2.1 Fisher's Information and Efficiency

Fisher's concepts of information and efficiency are especially used in the context of IRT. In that connection, the problems of *item information* in IRT models focus on scaling, test construction, and optimization. There are also some more recent concepts of information contrasting with Fisher's traditional statistical approach. Most of them are related to the early era of cybernetics and are usually based on the probability or empirical frequency of data. Shannon's information I_S from the early 1950s is the classic among those. It is based on the concept of probabilistic "uncertainty" measured by the distribution of probabilities/empirical frequencies over several categories of a research variable.

Let's consider categories of possible responses to a Likert-scaled attitude questionnaire: 1, 2, 3, 4, 5—from 1 "Fully disagree" up to 5 "Fully agree." We find the maximum uncertainty about a subject's score—now called *entropy*—when individuals of the given sample are distributed into the categories uniformly, in that example with equal proportions $p_1 = 0.20$, ..., $p_5 = 0.20$. We find the minimum uncertainty if all individuals were located in the same category, say $p_1 = 1.00$ with all other categories empty, i.e. if there is no dispersion, and $s^2 = 0$. In a further step, the concept of "information" is based on the intuitive idea of *reduction of uncertainty*, i.e. a decreasing difference of two entropy levels, before and after receiving a "message" from the diagnostic/measurement procedure of another variable, say x.

4.3.2.2 Shannon's Entropy

The measure of Shannon's entropy (uncertainty) $H(y)$ is the weighted mean of binary logarithms log_2 of the probabilities/frequencies of subjects in different alternative categories, say in the example of the Likert scale: p_1, p_2, p_3, p_4, p_5. Thus, $H(y) = \sum p_i log_2 p_i$ is expressed in a number of bits. After getting the "message" about the distribution of subjects on another variable (i.e., another Likert scale), used as an independent variable x, and about the conditional probabilities $p_i(y|x)$, we compute the decreased amount of the entropy of y given the knowledge of x, i.e. the conditional entropy $H(y|x)$. While Fisher's efficiency or "relative information" compared the original variance s_y^2 to the decreased conditional variance $s_{y|x}^2$ in terms of a ratio—Shannon's information I_S is expressed as the decrement in bits in terms of subtraction, meaning the original entropy minus conditional entropy: $I_x = H(y) - H(y|x)$.

Why is this abbreviated introduction to information theory presented here? There are two reasons:

1) The reader should be informed that there exist additional, different, and *alternative theories* and approaches to test reliability and validity that are not based on the traditional statistics of variance, correlation, and regression.
2) Further on, there is an interesting simple relationship between both conceptions of "information;" on one hand, the statistical Fisher's raw score information, efficiency, or "relative information" $1 - Ef$; and, on the other hand, the information-theoretic Shannon's I_S. Namely, under a bivariate normal distribution,[11] I_S is related to the correlation coefficient r_{xy}, shortly r, by $1 - r^2 = 4^{-I_S}$. Assume that y is a test–retest replication of x. Then, under this assumption,

11 Derived by Kullback (1959), applied to test theory by Blahuš (1969, 1971).

the CT theory unreliability, the Fisherian efficiency of measurement, and the information-theoretic Shannon's information I_S are mutually related by $Ef = 1 - Rel$ and $1 - Rel^2 = 4^{-I_S}$. In the end, this is an illustration supporting the opinion that different approaches may converge.

4.3.3 (§47) How to Recover the Unknown True τ? Randomness as the Fundamental Principle in Classical Test Theory

If we would like to follow the technological measurement approach in behavioral and social measurement, we would face the problem that we do not have either of the two procedures (1) and (2) (of Section §48) to fix the true value τ^* over a long run of n replicated measurements in one human subject. In some rare situations, especially in biomedical research and in some other branches where such technological thinking is sometimes applied, the gold standard method (1) can be used. It should yield a construction of approximation to τ^* with a precision of several orders higher than the ordinary methods, which approximation is then accepted and treated as if it were that would-like-to-be known true value. The main disadvantage of this approach is that usually it is not hundreds of times better than the calibrated field method, which distorts the function and formulas for constructing estimates of τ (in terms of a regression function).

We can compare laser measurement as a possible gold standard for the crude length measurement by a yardstick in physics, to some much weaker gold standards in the "soft" sciences. As an example, let us consider the measurement of body fat by the method of weighing a person once normally and a second time under water. The difference is then recomputed into the mass of fat in his/her body using the knowledge about the average specific mass of human fat (as found in cadavers). Then the field method of estimating body fat by measurement of the thickness of several skin folds is compared and correlated to the gold standard of the underwater weighing method. Thus, we can see that even the gold standard itself may not have excellent reliability. Even the procedure (2) working with τ^* as unknown but fixed – by keeping most important conditions constant to allow a longer run of diachronic replications – is not applicable as was discussed in Section §43. (Perhaps an exception could be somatometric measurement of certain bones, say the width of the pelvis as the distance between two well-identifiable anatomical points called anterior–superior spines.)

As we saw in Section §45, the main property of replications, which comes from the mathematical postulates of identical but independent distributions, treats replications as *random* in their full, general *probabilistic* sense. This *probabilistic randomness* produces special consequences, for instance:

- they are *a fortiori* independent in the sense of the linear regression, (i.e., uncorrelated) and
- their deviations around the mean converge to zero (recall $\overline{x} \longrightarrow \tau$), etc.

On the other hand, it often seems as if CT theory were trying to come near to the ideal of replications by using its specific concept of *test forms*. Their well-known special case is the "test–retest" tandem (only if the necessary corresponding conditions are met). Sometimes, instead of "test forms," some other terms are also used (i.e., "test measurements," etc.), but this term is confusing, possibly indicating that any instance of using a certain test to carry out measurement is its "form." Therefore, it is important to avoid confusion and insist that *test forms* must fulfill basic *statistical requirements* for their CTT model, namely *randomness* and *independence* formulated in terms of *regression* and *correlation* (otherwise, the use of the term is inappropriate in the sense of CT theory). By long historical tradition, these statistical requirements are called CT theory's "Axioms" which are introduced in detail in the next section. This means that randomness and independence for test

forms are not conceptualized in general probabilistic terms, but only through their special correlational variants (more specifically, expressed in terms of moments of variables up to the second order).

4.3.3.1 Parallel Forms

A very important special subcase of test forms is the case of *parallel forms*. Parallel forms are distinguished from general test forms by an additional statistical requirement, namely that the standard errors s_ε of the different test forms are mutually equal. It is important to be aware that the CTT concepts of *test forms* including *parallel forms* are statistical, *mathematical definitions* of a similar kind, as are the replications in the Gauss model. They are not intuitive conjectures of some modifications of the same test or its use under slightly different conditions that seem to be mutually very similar—by intuition "equivalent," "parallel," or so on. Even a high degree of agreement of several top experts who evaluated the semantic-content validity of several tests is not enough since the above-mentioned terms are formal-syntactic terms and mathematical definitions. These terms and formal concepts enable us to use CT theory in an analogic way as the formal concept of replications helps use the theory of measurement errors in the Gauss model. With its paradigm of only small n, typically $n = 2$ test forms measured over a sample of N individuals, we would not be able to approximate the individual person's errors ε_i, which might be possible had we used the gold standard method as an approximation of the true τ. Therefore, we are not able to compute their SD, namely the standard error s_ε, and the error variance s_ε^2, which would leave the definition of CTT reliability in section §45 by $Rel(x) = 1 - s_\varepsilon^2/s_y^2$ as purely hypothetical.

Fortunately, we have the theoretical and mathematical foundations consisting of the concepts of test forms and parallel forms in CT Theory at our disposal. Only a very superficial view would lead to a conclusion that either behavioral modelers or mathematicians just willfully create assumptions about randomness of errors and/or that practical experience suggests that errors usually behave as more or less random. Seen in a wider philosophical context, the foundation comes from the gnoseological (theory of knowledge) opinion that we can get our knowledge through its contrast against a certain background—here the background of "perfect randomness." By the empirical conditions of measurement in the behavioral sciences, and by the conceptual tradition of CTT, randomness is formulated in terms of CTT axioms. However, the philosophical foundations of randomness are the same in the theory of measurement errors as typically expressed in the Gauss model (as well as in CT theory and its model). While the randomness is produced by the two probabilistic postulates/axioms in the former case, it is specified by CTT axioms about regression and uncorrelatedness in the latter case. If the CTT axioms on randomness are confronted with the probabilistic concept of replications, they manifest themselves as just direct consequences of the former.

4.3.3.2 An Application of CT Theory

Let us see an artificial example as an application-oriented illustration of the character of CT theory. In the case of Gaussian measurement errors, we had a model whose measurement function Φ was a stochastic identity, whereas now in CT theory, the general stochastic function with its probabilistic postulates is specialized into a regression measurement function Φ,and a *regression identity*. Due to the typical number of $n = 2$ test forms, we actually have two such regression models: $x_1 = \tau + \varepsilon_1$ and $x_2 = \tau + \varepsilon_2$, in which τ is assumed to be a constant over the $j = 1, 2$ observations, but varying over the $i = 1, 2, ..., N$ individual subjects. Therefore, if considered for values in individual subjects, we have two sets of equations for the values: $x_{1i} = \tau_i + \varepsilon_{1i}$ and $x_{2i} = \tau_i + \varepsilon_{2i}$, where $i = 1, 2, ..., N$, and τ_i is the ith subject's unknown constant, and ε_{1i}, ε_{2i} are his/her unknown errors on the two parallel forms.

Table 4.2 An artificial illustration of parallel forms of the same test-specific true scores: fictive scores of verbal fluency test [number of words in 1 minute].

Subject *i*	1	2	3	4	...	*N*	Expectation	Variance
Unknown true score τ_i	10	14	6	9	...		$E(\tau) = 14$	$s^2_\tau = 6$
First parallel test form								
Observed scores x_{1_i}	12	15	5	12	...		$E(x_1) = 14$	$s^2_{x_1} = 10$
Errors $\varepsilon_{1_i} = x_{1_i} - \tau_i$	2	1	−1	3	...			$s^2_{\varepsilon_1} = 4$
Second parallel test form								
Observed scores x_{2_i}	9	14	8	7	...		$E(x_2) = 14$	$s^2_{x_2} = 10$
Errors $\varepsilon_{2_i} = x_{2_i} - \tau_i$	−1	0	2	−2	...			$s^2_{\varepsilon_2} = 4$

Note also that numerically: $s^2_{x_1} = s^2_\tau + s^2_{\varepsilon_1} = s^2_{x_2} = s^2_\tau + s^2_{\varepsilon_2}$.

As was mentioned above, the special concept of parallel test forms assumes that both test forms have the same standard errors, here it means that $s_{\varepsilon_1} = s_{\varepsilon_2} = s_\varepsilon$, or in the form of variances $s^2_{\varepsilon_1} = s^2_{\varepsilon_2} = s^2_\varepsilon$. The illustration is in Table 4.2. The numbers in Table 4.2 are artificial—of course, the true values and errors are unknown in reality. Nevertheless, for the sake of illustrating the statistical characteristics, the expected means $E(.)$, and variances $s^2_{(.)}$, have been intentionally issued so that they correspond to the properties of CTT postulated in the next section. For example, the two means $E(.)$ are equal, as well as the two observed test and retest variances, and the sum of error variance and true variance gives the observed variance.

4.3.3.3 The Expectation Principle

There are two theoretical and mathematical conceptions under which the parallel forms approach to the principle of randomness may be applied, and the CTT model derived. While each of the conceptions treats the same principle—randomness—in a different way, both result in identical outcomes with respect to the fundamental theorems and formulas, for example, the very important formula for approximate computation of the reliability coefficient. These two conceptions are:

1) the well-known older conception of *Platonic* true scores and
2) the e*xpectation conception*, in which the true score is *defined* as an *expectation* value.

In the framework of the Platonic conception, the true score τ is assumed as a certain exact value that is *inaccessible in principle* to the researcher. This unknown τ exists with definite certainty out in the world of ideas, beyond the reach of researchers – who are deep within Plato's cave, where all we can see are the shadows on the cave walls (i.e., the observable scores x). This standpoint is a kind of epistemic skepticism about the human capacity to get complete knowledge of the world, even in an infinitely long run of human attempts to apply such an effort. The true knowledge definitely lies beyond the horizon of such an effort.

On the contrary, according to the expectation conception, the true τ is accessible at least through theoretic operations. Namely, we may follow the idea of more and more replications as in the Gauss model, and the problem becomes rather epistemological and depends on the researcher's ability to ensure the process of replications properly. In this variant of CT theory, τ is theoretically accessible through the hypothetical operational process by an infinite number of parallel forms. To illustrate the situation by analogy with replications of x in the Gauss model, we can say that now,

surprisingly, τ is *not estimated* by the construction function Φ (in the form of the mean of x, "$\overline{x}$", which converges, $\overline{x} \longrightarrow \tau$ as $n \longrightarrow \infty$, to the unknown τ), but, a little strangely, τ *is defined* as *that mean* (though it is unknown). In terms of CT theory, the expectation variant of the true score is given by the *expectation definition*, $\tau \equiv E(x)$, with $n = \infty$ parallel forms. Historically, a similar concept used to be known as the score of an "infinitely long test" when the rule "the longer the test, the higher the reliability" comes to limit. However, the possibility that this definition is not only a wishful claim—or even a willful assertion by a family of scientists with extremely high self-esteem—can be supported only by a hope that due to the infinitely long series of instances, the randomness will become prevailing and the random errors would eliminate each other. Hence, it expresses an epistemic hope that exact knowledge is accessible to us in principle, at the end of an infinite process just beyond the horizon.

4.3.3.4 The Randomness Principle

It is interesting to see how both of the conceptions are implicitly based on the same philosophical and general methodological principle of randomness. However, this fact has not been explicitly expressed in the original works on this approach and is not widely known. The old and traditional Platonic conception presents its foundation decomposed into three or four axioms of CT Theory. But, in CT theory, it usually stays hidden that the set of its axioms is nothing more than a consequence of the general principle of randomness presented in its three or four separate aspects.

On the other hand, the more recent expectation conception begins with a rather synthetic and simple starting point, the expectation definition. This is especially plausible from the mathematical point of view since it shows that the axioms of the older Platonic conception are simply consequents of this new definition of the true score in the expectation conception. Or, which is the same, that all fundamental theorems including reliability, etc., follow directly from this new definition and, therefore, the axioms of the Platonic true score approach are not necessary. Nevertheless, formal simplicity alone should not be the only justification for choosing any model or approach. Even the fact that the expectation definition eliminates certain mathematical controversy, which the Platonic conception faces in the case of dichotomous scoring (zero/one item), is rather formal and deals just with this special way of scoring. It would seem that the reasons should be based more on the conceptual aspects.

Hence, both conceptions of the specific true score are based on a purely formal assumption that is represented mathematically either by (i) a set of four axioms (described in Section §48) and/or by (ii) just one general axiom (i.e., the expectation definition). From the point of view of general methodology, the difference between the two is mostly formalistic, consisting of different levels of parsimony. However, from the point of view of the philosophy of science, it seems interesting how the randomness principle, on which both of them are founded, enables our access to knowledge. The complete uncertainty in the wide background enables us to acquire more certainty in a narrower area. We hope that the ideally complete randomness in the surroundings of τ allows us to elaborate a method (replication) to get information about its better estimate. It poses randomness in the role of chaos against which we are able to see the regularities. Isn't it similar to measuring radiation against its background? The same principle works in randomized experiments. It may seem, a little strangely, that randomness acts as a fundamental epistemic postulate and a desired factor in the frame of scientific inquiry.

In the end, let us imagine theoretically what would happen if the two definitions of the true score τ were applied to the measurement of errors in Gauss model and its replications. In the first case with the CTT axioms, the new special case, possibly called the "Gaussian model for parallel forms," would become a linear correlational specialization of the original Gauss model and its replications. The correlational independence would be substituted for the original probabilistic one. In the second case, if the unknown true value τ^* were *defined* by $E(x)$ of the replications x in the Gauss model,

the simplifying consequences might be interesting—not to mention the conditions that are otherwise necessary for application of the central limit theorem that ensures the convergence to τ^*.

4.3.4 (§48) Reliability of τ in Terms of the Platonic Conception of CTT

There are four axioms A1–A4 of the older Platonic conception mentioned above. They are presented below with an indication of how they follow conceptually from the intuitive general assumptions of randomness and independence. However, first, some additional postulates have to be made. For several forthcoming derivations and examples, we will need more than just $n = 2$ test forms x_1, x_2 as in Table 4.1. For our purpose, it is necessary to work with four forms, and four separate subtests to later show the difference between the specific and generic scores. The resulting situation may be illustrated by adding further rows of parallel forms, scores, means, and variances into Table 4.1. Then, we measure the unknown but *the same* τ, constant over different test forms x_j, x_k, ... in each of N subjects, by assuming the CT theory model with the stochastic identity in the form of special regression measurement functions $\Phi: x_j = \tau + \varepsilon_j$, $x_k = \tau + \varepsilon_k$ with the test forms j, k... = 1, 2, ..., n, and $n \geq 4$.

In addition, it is quite reasonable that any diagnostic method should not have zero variance in order to be able to distinguish between people. Further, its real-world scale or ruler should not be supposed to have infinite length. Therefore, we assume "reasonable" variability over the N subjects: Nonzero and finite variances s_j^2, s_k^2, s_τ^2, it means $0 < s^2 < \infty$ for any of those. Having these preliminary postulates, we formulate the so-called axioms.

4.3.4.1 Four Axioms of Classical Test Theory

1) Since the parallel forms differ from each other randomly, then by common sense, their deviations, the errors ε, around the values τ should be distributed symmetrically, i.e. without a systematic shift, hence the mean (expectation) of these errors should be zero. This could be called *the axiom of counter-balanced errors*: A1 $E(\varepsilon_j) = 0$; $E(\varepsilon_k) = 0$ for each of the n test forms.
2) Due to their perfect randomness, the errors should not be related to anything that is somehow systematically related to the diagnostic/measurement procedure. (Of course, the true levels of τ are supposed to be related to the observed diagnostic outcomes, the values of x, since the diagnostic procedure x has been intentionally constructed with a belief in a nonzero validity with the correlation $r_{x\tau} \neq 0$). This assumption can be expressed as *the axiom of uncorrelatedness of errors with the true values*: A2 $r_{\varepsilon_j \tau} = 0$, $r_{\varepsilon_k \tau} = 0$ for each of the n test forms.
3) By the ideal of a completely random mixture of the errors in their whole noise, even the errors themselves should not be interrelated in any way, which could be represented by *the axiom of uncorrelated errors*: A3 $r_{\varepsilon_j \varepsilon_k} = 0$ for each pair of different test forms, i.e. for $j \neq k$.
4) Since the conditions and the environment are assumed constant over the $j, k = 1, 2, ..., n$ test forms, the "amount of noise" should not be different in the parallel test forms. Hence, the dispersion of the errors over the N individuals in the separate forms should stay the same. This expresses statistically what is meant by "the real world to be constant"—stable up to chance oscillations. This assumption of constant error variance, hence constant SD of ε, is the axiom of equal standard errors, or *the axiom of classical parallelism*: A4 $s_{\varepsilon_j}^2 = s_{\varepsilon_k}^2$ for any pair of n *parallel* test forms j, k.

The different parallel forms of a test—and/or different parallel tests—are then also referred to as "*classically parallel*" or "strictly" parallel. Since such an exact form is not found in reality, it is also said that the tests are assumed to be "nominally parallel" or "essentially parallel." It means that in the ideal case, any of the n parallel forms x_j, taken separately, should be as good for estimating τ as any other of the rest of the $n - 1$ ones. Therefore, the axiom A4 of parallelism expresses the highest

level of mutual *equivalence* or interchangeability—whichever parallel tests you use, the information, error and so on are the same.

4.3.4.2 Using Classical Test Theory

Now, we are almost ready to use A1–A4 for the derivation of the whole arsenal of well-known traditional theorems of CT theory, the powerful formalistic foundations of measurement in the behavioral sciences. These were laid at the beginning of the 20th century by Spearman and Brown, more fully explained in Gulliksen's textbook in the 1950s (Gulliksen, 1950), finally becoming a classic after they were embodied into a more modern and general approach in the Statistical Theories of Mental Test Scores by Lord and Novick (1968). Nevertheless, we are going to present them in just a narrow selection focused on reliability as related to the concept of precision of measurement.

First, we intend to make the following derivation as simple as possible. Therefore, we will assume that the measurement scale has been shifted so that its zero point is at the mean, in the same way as introduced in Appendix 2 and Chapters 1 and 2, and we will work with the *centered* or *deviation scores*, ${}_dx \equiv (x_i - \overline{x})$. Second, we should recall three auxiliary formulas: (i) the definition of covariance and variance from Appendix 2 rewritten here into deviation form, (ii) the relationship of s_{xy} and r_{xy} from Appendix 2, and finally, (iii) the well-known and easily derivable[12] formula for the variance of a sum of variables:

1) $s_{xy} = (1/N)\sum(x_i - \overline{x})(y_i - \overline{y}) = E\left({}_dx_dy\right)$, $s_x^2 = (1/N)\sum(x_i - \overline{x})^2 = E\left({}_dx^2\right)$,
2) $s_{xy} = r_{xy}s_xs_y$,
3) if $u = x + y$ then $s_u^2 = s_x^2 + s_y^2 + 2s_{xy}$.

Then we can proceed:

1) By the very definition of the CTT model, if the observed scores are on centered scales, i.e. in deviation form ${}_dx$, for all of the n test forms x_j, x_k, ..., then, since the errors are in deviation form by axiom A1, the same holds for the true scores $\tau \equiv {}_d\tau$. Furthermore, it means that $E(\tau) = E(x)$, namely that the mean of the observed score is unbiased with respect to the true score mean. Further, for example, it also follows that $s_\tau^2 = E(\tau^2)$.
2) Since the CTT model presents ${}_dx$ as a sum of their true scores and errors, then application of the formula (*iii*) gives $s_x^2 = s_\tau^2 + s_\varepsilon^2 + 2s_{\tau\varepsilon}$, but by A2 and formula (2), it reduces to the simple additive decomposition of the observed variance into the true score and error variances: $s_x^2 = s_\tau^2 + s_\varepsilon^2$. In addition, a consequence of the axiom A4 is that using different parallel forms x_j, x_k, ..., we may approximate the same τ. Also, as it results in the equality of the known variances $s_j{}^2$, $s_k{}^2$, ... we may roughly check how well the axiomatic assumptions hold in reality.
3) The crucial theorem comes out by application of the axioms to the *covariance* s_{xy} *of two test forms* x_j, x_k, represented by ${}_dx_j = \tau + \varepsilon_j$ and ${}_dx_k = \tau + \varepsilon_k$. From the standard definition as in (*i*), $s_{jk} = E({}_dx_dy)$, we get by substitution

$$\begin{aligned} s_{jk} &= E\left[\left(\tau + \varepsilon_j\right)\left(\tau + \varepsilon_k\right)\right] \\ &= E\left[\tau^2 + \tau\varepsilon_j + \tau\varepsilon_k + \varepsilon_j\varepsilon_k\right] \\ &= E(\tau^2) + E\left(\tau\varepsilon_j\right) + E(\tau\varepsilon_k) + E\left(\varepsilon_j\varepsilon_k\right), \end{aligned}$$

12 If we consider all three variables in deviation form, then:

$$s_u^2 = (1/N)\sum {}_du^2 = (1/N)\left({}_dx + {}_dy\right)^2 = (1/N)\left({}_dx^2 + {}_dy^2 + 2{}_dx_dy\right) = s_x^2 + s_y^2 + 2s_{xy}$$

which due to uncorrelatedness by A2 and A3 with the use of formula (2) becomes the *CTT fundamental theorem*: $s_{jk} = s_\tau^2$.

I will call this the *fundamental theorem of CTT* since most further theorems and derived formulas depend on it. As far as the axioms hold, the theorem enables one to approximate the *unknown true variances* of the true scores in terms of the *known covariance of parallel forms*, and practically compute

$$Rel(x) = s_\tau^2/s_x^2, Unrel(x) = s_\varepsilon^2/s_x^2.$$

Note that A4 is also needed to assume the *same* unknown true variance. In terms of the other conceptions, it also enables us to compute the measurement imprecision by *Imprec* $(x) = s_\varepsilon/s_x$, efficiency $Ef(x)$, and relative information $1 - Ef(x)$ introduced in sections §50 and §51.

4.3.4.3 Some Practical Applications of Classical Test Theory

It is very important to be aware that the CTT fundamental theorem may work in practice only *as far as the axioms are met* by real-world conditions in the research practice of testing or measurement. Therefore, reliability may be only *approximated*, never "established" or "estimated." If sometimes one can read that "reliability has been estimated by the test-retest method, this "estimation" is to be understood in the sense of natural or layman's language where any approximation or guessing is often called "estimation." Namely, the term "*estimation*" might be confused with the notion of estimation in inferential statistics, a regression estimate, or so on. As regards further aspects of "estimating reliability," we should be aware that in a standardization/calibration study, the reliability coefficient of a test in a certain population can be statistically estimated with its confidence interval on the basis of the reliability coefficient, and approximated in a random sample drawn from that population.

The fundamental theorem has the most practical consequences for approximating the reliability $Rel(x)$ of test x. By definition, the correlation coefficient of two test forms is $r_{jk} = s_{jk}/s_j s_k$. If the two forms x_j, x_k of the test x may be assumed classically parallel, then from A1 and A4 we have $s_j = s_k$, which should hold in empirical findings, at least approximately. Then $s_j s_k = s_x^2$ and by the CTT fundamental theorem, we get $r_{jk} = s_\tau^2/s_x^2$, the well-known theorem: "Correlation of parallel forms is equal to reliability."[13] In practice, it holds as far as the forms are close to the ideal of parallelism. If the practical approximation of the parallel forms is based on the "*test–retest*" *method*, then sometimes a special notation is used by some authors. They do not rarely use the prime for the retest parallel form: say x, respectively, x_j for the first session parallel form, and x', respectively x'_j for

13 Warning: Once having been approximated by the computational procedure for $r_{xx'}$, the reliability $Rel(x)$ is NOT to be interpreted as a correlation coefficient anymore. That computational possibility should be considered only as a coincidence. Therefore, it would be nonsense to compute its square in a mistaken attempt to calculate a coefficient of determination, and interpret it falsely as a proportion of "explained" or "predicted" variance of the retest x_2 by knowledge of scores in the first testing x_1. The reliability $Rel(x)$ is already a proportion of the "explained" variance, viz. variance of true score τ with respect to the observed variance of x. Reliability does not deal with predicting the post-test from pre-test—it may be an interesting problem in a longitudinal predictive study based on using the same test. Note the different terminologies:

a) test and retest as two parallel forms of a test in the reliability study by repetition, vs.
b) pre-test and post-test in a longitudinal study or intra-group experiment focused on increments in time.

We would not have presented this warning if we had not found that erroneous explanation in a book of a well-known statistician and internationally recognized scholar who, unfortunately, was not a specialist in test theory or behavioral statistics. An example of this mistake may be found in Wang (1993, pp. 132–133).

the retest session parallel form, and $r_{xx'}$ or $r_{jj'}$ for a computed correlation coefficient that approximates the reliability $Rel(x) \approx r_{xx'}$.

If we know the typical value of $Rel(x)$ in a certain kind of population, as well as its typical standard deviation s_x, we can compute an approximation of s_ε, the *standard error* due to unreliability, in the measurement units by $s_\varepsilon = s_x\sqrt{1 - Rel(x)}$. For instance, as we have presented before, the V-fluency test has standard deviation $s_x = 4$ (words) in the sample, and we assume that its reliability had been found $Rel(x) = 0.91$. Then we could compute its standard error: $s_\varepsilon = 4\sqrt{1 - 0.91} = 4\sqrt{0.09} = 4 \times 0.3 = 1.2$(words). Under the assumption of a normal distribution, this magnitude of that "average" error means that the true score lies in the interval ± 1.2 around our observed score with probability of 68% (cf. section §18). One may consider either the practical, clinical acceptability of this "average" error and/or its scientific importance. But, in addition, it is sometimes useful to think about the standard error in z-scores, i.e. in SDs, and to realize that this $s(z)_\varepsilon$ is simply $s(z)_\varepsilon = \sqrt{1 - Rel(x)}$ [SD], cf. the relationships to precision/imprecision (Sections §48 and §50). In this example, it means that the standard error is ± 0.3 (SD), i.e. about one-third of a standard deviation.

4.3.4.4 Problems in Applying CT Theory to the Reliability of Derived Units

We are considering here potential problems in creating composite scores from derived items. Derived units were introduced in Section §42. Consider the example of Body Mass Index (BMI), specifically $y = x_1/x_2^2$. It has the following interpretation: BMI = (body mass in kg)/(body height in m). In connection with CT theory, it may seem interesting to look at the problem of reliability of this nonlinearly constructed index. On one hand, each of the independent variables x_1, x_2 can have a real-world counterpart of its own CTT parallel form $x_1 = \tau_1 + \varepsilon_1$, and/or $x_2 = \tau_2 + \varepsilon_2$. On the other hand, in the CTT model of the constructed dependent variable $y = \tau_y + \varepsilon_y$, the CTT axioms are broken due to the nonlinear deterministic function, which certainly produces nonzero linear correlations.

If we are interested in a CTT model for BMI, i.e. for y where $y = x_1/x_2^2$, then we get $y = \tau_y + \varepsilon_y$, where $\tau_y = \tau_1/(\tau_2 + \varepsilon_2)^2$, and $\varepsilon_y = \varepsilon_1/(\tau_2 + \varepsilon_2)^2$. This clearly shows that the errors ε_y are not independent of τ_y but are functionally related to it. Unfortunately, many other similar indices are constructed just by intuition and with an oversimplified view, e.g. in biomedical studies, oxygen consumption is related to a kilogram of body weight and so on (although some more rational methods are easy to construct).

Such examples should show how problematic it can be to use a CTT model for approximating reliability in measuring a derived unit—usually reliability of a "composite score" constructed as a nonlinear "formula score." This holds even in the case when the independent variables (subtests) can be plausibly modeled by CTT separately and their parallel forms can be very well approximated in reality, say by test–retest forms. Then, in general, the nonlinearly constructed dependent variable still breaks the CTT axioms. Hence, for example, the "test–retest" method cannot be assumed to yield an approximation of reliability when applied to such a nonlinearly constructed composite.

4.4 Summary

This chapter has substantial additions from recent work, resulting in some welcome simplifications and deletions. Both *Psychometrics* and *Kinanthropometrics* are disciplines characterized by measurement and measurement instruments. This makes them conceptually part of International

Metrology, even though this kinship is only recently becoming a real possibility. Metrology, the science of measurement, has Applied and Legal aspects that arise from good science. This results in excellent applied measurement instruments and public trust in them.

A draft of Chapter 8 on International Metrology for this book was initiated in 2019, and Petr's incomplete ideas about Ostensive Definition and Ostensive Demonstration were evolving into what we have made of them in this volume. The biggest impact of recent work on this chapter has been the book *Measurement Across the Sciences* (Mari et al., 2021). The difference this book can make on all Metrology, including especially the human sciences, evolves from the fact that Mari is an Italian engineer-metrologist. He sat on the committees that released the Vocabulary International of Metrology volume 3 (VIM3) in 2012, and he and his coauthors phrased the new definition, starting with (and augmenting) VIM's language. Thus, it has a chance of being added to future volumes of VIM. It offers the hope for human sciences and all sciences that versions of the 63 concepts that are new or revised from what is in VIM will be accepted. This will open the doors of international metrology to "All Sciences and Technology, at any level of uncertainty." This international definition of Metrology, the science of measurement, is a hopeful aspiration because the VIM, as currently written, excludes the possibility of interval measurement as it has evolved in human and life sciences.

The impact of the new definitions on this chapter has been liberating, simplifying, and powerful:

1) Direct measurement has been defined as involving an instrument in the definition, and values on a scale defined so that ratio, interval, ordinal, and even categorical levels of measurement can be included; each level must reach different but high standards.
2) Indirect measurement involves two or more direct measurements and a calculation, making indirect measurement a subset of derived measurement.
3) The definition of direct measurement includes three functions that move an empirical observation from the empirical world into the instrument scale, and then to a public scale in the information world. These fulfill Petr's search for associative functions between the empirical world and the world where explanations of the scale values must be made.

These simplifications replace the structure Petr was developing using a concept called Associative Measurement. This concept was not endorsed by the successor authors nor by newer and older writers. Mari et al. (2021) offered a better alternative. Earlier, Luce, Krantz, Suppes, and Tversky (1990) regarded it as a variation on conjoint measurement that they did not explore. But the new definition of direct measurement accomplishes the "associative functions" Petr sought with much more clarity and power so that any scientist who comes to understand it will be delighted with how much more it accomplishes, and how much greater simplicity it brings. Confusion and circularity have surrounded definitions of direct, indirect, and derived measurement for many years before the Mari et al. analysis and proposals.

One other aspect of this new definition brings in another very significant impact on this chapter. It is that "measurement is an empirical and informational process *designed on purpose*." A discussion of instrument design and development has been side-stepped in the definition of validity, something we criticized in Chapter 3. In Chapter 4, we addressed this gap by references to instrument design and development. The references in Section 4.2 on Evidence-Centered Design, lead erstwhile instrument developers to guidance that also address Petr's oft-expressed concerns about "so-called constructs." Scientists in the human domains must apply much more discipline to instrument development than has been the custom in the "*'common law' of test development conveyed through advice and opinion through its history*"—a comment by Cynthia Board Schmeiser of ACT—whose college admission tests are carefully and systematically developed.

The Gauss measurement-errors model (see Section §41) serves to link observed measurements with their corresponding true values recognizing the inevitability of random errors of measurement. These measurement errors can be described in terms of uncertainty across multiple replications of a common measurement procedure. As identical "clones" of the measurement procedure, replications have both (i) a semantic-content interpretation and (ii) a formal, mathematical meaning. In the Gauss model, replications are considered to be (i) probabilistically independent, and to have (ii) the same probability distributions. Ultimately, replication is recognized as an idealized mathematical concept to be sought and approximated, but unlikely to be realized in actual measurement situations.

Efforts are sometimes made to approximate replications through repeated measurements under conditions that are believed to remain constant (even if only momentarily) over time. This represents the diachronic approximation of replications. Other times we may try to approximate replications by simultaneously obtaining multiple readings from appropriately dispersed measurement "sensors" intended to reflect a shared true value. This is referred to as the synchronic approximation of replications. Alternatively, it is possible to think of replications as randomly drawn sample realizations from an appropriate (sub)population using a standard inferential statistical framework.

In LCF models, alternative continuous indicators of the same underlying attribute may utilize different measurement metrics, potentially varying in their origin or zero point or in their unit of measurement, or both. This is called a congeneric measurement model[14], meaning that the indicators are generated from a common source (common factor) without other restrictions on the item parameters. Indicators can also vary in their sensitivity or ability to detect small differences in the latent trait. Indicators of a common attribute can have both shared and unique components. In general, without other information, it is difficult to determine how much of an indicator's unique component is specific to the indicator and how much is simply due to random error. An indicator's nonerror or overall true score component is comprised of a generic true score component that is shared with the other indicators of the construct, and a specific true score component that is unique to the given indicator but is also not random error.

In more extensively studied areas, there can sometimes be a so-called gold standard measure (see Section §44) that serves as a reference for subsequent or alternate measures of the designated property or attribute. The quality of a new measure is judged in terms of its deviations from the measurements using the gold standard. Whereas the Gauss model way of thinking utilizes the notion of relative error or (im)precision, CTT utilizes the idea of (un)reliability. In practice, the two ideas can be shown to be equivalent and also directly related to Fisher's formulation of information efficiency. Without a specific gold standard, CTT uses the idea of equivalent or parallel forms to help us assess the reliability of measurements.

The concept of probabilistic randomness along with the expectation principle from mathematics and statistics provides a particularly useful and more general way to express basic principles of the Gaussian theory of measurement errors and CTT. These underlie particularly the development of four main axioms of CTT, (i) error counterbalancing, (ii) uncorrelatedness of the errors with the true values, (iii) mutual uncorrelatedness of the errors, and (iv) equal errors for parallel tests. These can then be used to produce a broad range of theorems and corollaries in CTT (see Lord & Novick, 1968, Chapter 3).

14 See *Computational Guide 1, Classical Test Theory: A Model-Based Approach*.

4.5 Study Questions

A. Discussion Questions

4.1 Introduction and Overview of "Unfinished Business" from Chapter 3

Q1 What is a major goal of domain theory as explained at the beginning of paragraphs labeled as "domain theory"?

Q2 What is meant by "learner-centered assessment" and what does it mean for the future of educational and psychological testing?

Q3 In what ways could the growing focus on learner-centered assessment be of benefit to students as they prepare for a career? How could it be of benefit to the workplace?

Q4 What are some of the ways you would expect that psychological and educational testing methods would be found to differ across a variety of domains?

Q5 The full page titled "Assessment in relation to measurement" in Section 4.1 to Chapter 4 begins with these two sentences, "Assessment is a word from human measurement activities. It includes but is more than measurement instrument development and use." Based on what you learn from this page, write a brief and perhaps visionary account of how *human measurement activities,* at their best, have the potential to greatly enrich our individual human lives.

Q6 What kinds of changes and growth can be expected in the sciences of psychological and educational testing as a result of the recent advances in general science of metrology?

§41 The Gauss Model for Replications

Q7 What are "replications"? What do the authors mean when they say, "In summary, true replications are generally impossible to achieve in practice ... If the errors are random, approximately replicated measurements can help us obtain an estimate of the true value and an estimate of the error variability."

Q8 Explain the importance of the "randomness of errors in the replications ..." as a precondition for the Gauss model to be valid.

§42 On *Direct*, *Indirect*, and *Derived* Measurement Across the Sciences

Q9 Explain the three main links between the empirical world and the information world defined by Mari et al. (2021).

Q10 Define direct measurement, derived measurement, and indirect measurement.

Q11 What do you suppose the authors mean when they ask "how does direct measurement make indirect measurement possible?"

Q12 Explain the authors' statement that "Directly measuring objects in classical ways is not the standard and most accurate form of measurement in today's world." What examples do they give to demonstrate their point?

Q13 Describe with examples how "coherent derived units in the SI (can be) expressed in terms of base units."

§43 The Specific True Score of a Separate Subtest/Indicator

Q14 From the final paragraph of Section §43, explain the three variables of classical test theory (CTT), the properties of each, and how they relate to one another.

Q15 Explain the fundamental rationale of "true score theory" and how true scores can be estimated.

§44 The Test as a Measuring Device, Compared to the Precision of Measurement in Technology

Q16 From the first paragraph of Section §44, explain the difference between the *standard error of measurement* s_e (as explained in Chapter 1, section §13, and the *standard error of regression estimate* $s_e = s_{y|x}$ (Appendix 4)).

Q17 From footnote 7, explain how the meter's scientific definition changed in 1983 and 2019, and how the meter is measured today.

§45 Reliability vs. Precision

Q18 Describe the relationship between *imprecision* from the perspective of metrology and unreliability from the perspective of classical test theory (CTT).

§46 Reliability of a Test, Efficiency of Measurement, and the Information Gained

Q19 Compare Fisher's concepts of *information and efficiency* with Shannon's measure of *entropy*. How are they similar and how are they different?

Q20 Describe which measure, Fisher's or Shannon's, is more closely associated with IRT (which stands for "Item Response Theory," as explained in §53 of Chapter 5). Please also give the rationale for your answer, that is, why it is more closely associated with IRT?

§47 How to Recover the True τ? Randomness as a Principle in Classical Test Theory

Q21 What is meant by "parallel forms" of a test?

Q22 Why do the authors emphasize the importance of *probabilistic randomness* and *independence of errors* as the basis for using parallel test forms as the theoretical and mathematical foundation in the development of classical test theory (CTT)?

§48 Reliability of the Specific True Score τ—In Terms of the "Platonic Conception" of CTT

Q23 What is "the Platonic conception of CTT" and how does it relate to the reliability of τ (tau)?

Q24 Summarize and explain the four axioms of classical test theory (CTT).

Q25 Explain how the principles of *mathematical expectation* and *randomness* provide a grounding for deriving the four fundamental axioms of CTT.

Q26 Summarize the *fundamental theorem of classical test theory* (CTT). That is, explain how, if you have parallel forms of a test and all four CTT axioms are met by real-world conditions for the test occasion, the theorem can be used to obtain a reasonable estimate of reliability for that test, $Rel(x) = s_\tau^2/s_x^2$.

B. Exercise Question

EQ1 The following Stata syntax simulates two strictly parallel tests following the description given in Table 4.1 of section §42. The preliminary commands clear the workspace, assign the square root of the true score variance (the standard deviation) to a local macro, and set a replication seed for the random number generator. In the next section, the drawnorm command is used to draw values from a normal distribution with a specified mean and standard deviation. A normally distributed true score with a mean of 14 and a standard deviation of $\sqrt{6}$, and two error scores with means of zero and standard deviations of 2 are simulated, and two observed scores are calculated by combining the true and error scores. The means, standard deviations, and correlations of the true, error, and observed scores are then displayed. Finally, Stata's sem command is used to estimate the three model parameters (true score variance, error variance, and mean) based on only the simulated observed scores.

```
*** Simulation of (strictly) parallel tests
*** Classical Test Theory (CTT)
*** Chapter 4, Section 42, Table 4.1
clear                       // Clear workspace
local s : display sqrt(6) // True score standard deviation
set seed 54321              // Set seed for replication

*** Draw true score and two errors (N=10000)
drawnorm t e1 e2, n(10000) mean(14 0 0) sd(`s' 2 2)
generate x1 = t+e1          // Observed score test1
generate x2 = t+e2          // Observed score test2

correlate t x1 x2 e1 e2, /// Correlations
 means                   //  Means

*** Structural equation model for two parallel tests
sem (T -> x1@1 x2@1)      /// Fix loadings
 (_cons -> x1@m x2@m),    /// Equate means
 variance(e.x1@e e.x2@e) //  Equate residual variances
```

Means and Standard Deviations

Variable	Mean	Std. Dev.	Min	Max
t	14.03907	2.448842	5.475672	23.41587
x1	14.02572	3.151452	1.05833	25.9879
x2	14.03032	3.175804	1.215784	26.04932
e1	−0.0133514	2.018543	−8.378093	7.682756
e2	−0.0087553	2.003328	−7.443583	7.934159

The true score and observed score means are approximately equal to 14, and the means of the error scores are approximately equal to 0. The standard deviation of the true score is approximately equal to the square root of 6, the standard deviations of errors are approximately equal to the square root of 4, and the standard deviations of the observed scores are approximately equal to the square root of 10, the sum of the true score and error variances (10 = 6 + 4). These values correspond to the results expected from CTT.

Correlations

	t	x1	x2	e1	e2
t	1.0000				
x1	0.7680	1.0000			
x2	0.7760	0.5949	1.0000		
e1	−0.0141	0.6295	−0.0126	1.0000	
e2	0.0077	0.0042	0.6367	−0.0028	1.0000

The correlations between the true and error scores are approximately zero, as is the correlation between the two observed scores. The correlation between the two observed scores is approximately

0.6, and the correlation of the true score with the observed scores is approximately $\sqrt{0.6}$. These again correspond to the expectations of CTT.

The Stata sem command analyzes only the simulated observed scores, using structural equation modeling to recover the true score information by treating the true scores as a latent variable. The model for strictly parallel tests is estimated by fixing the factor loadings to 1, equating the observed score means, and equating the observed score residual variances.

	OIM					
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Measurement						
x1						
T	1	(constrained)				
_cons	14.02802	0.0282494	496.58	0.000	13.97265	14.08339
x2						
T	1	(constrained)				
_cons	14.02802	0.0282494	496.58	0.000	13.97265	14.08339
var(e.x1)	4.054758	0.0573429			3.943912	4.168721
var(e.x2)	4.054758	0.0573429			3.943912	4.168721
var(T)	5.952935	0.1164437			5.729029	6.185592

LR test of model vs. saturated: chi2(2) = 0.94, Prob > chi2 = 0.6240

This provides the observed score means of approximately 14, the error variances of approximately 4, and the true score variance of approximately 6.

References

Berka, K. (1983). *Measurement. Its concepts, theories, and problems*. Dordrecht: Reidel.

Downing, S.M. and Haladyna, T.M. (ed.) (2006). *Handbook of test development*, firste. Routledge. ISBN: 9780203874776.

EAL (1997). *Expression of the uncertainty of measurement in calibration EAL-R2*. European Cooperation for Accreditation of Laboratories (EAL), West European Calibration Cooperation WECC 19-1990.

Ellis, B. (1966). *Basic concepts of measurement*. Cambridge University Press.

Federer, W.T. (1973). *Statistics and society: Data collection and interpretation*. New York: Marcel Dekker.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Krantz, D.H., Luce, R.D., Suppes, P., and Tversky, A. (1971). *Foundations of measurement: Volume I. Additive and polynomial representations*. New York: Academic Press.

Lane, S., Raymond, M.R., and Haladyna, T.M. (ed.) (2016). *Handbook of test development*, vol. *2*. Routledge.

Lord, F.M. and Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luce, R.D., Krantz, D.H., Suppes, P., and Tversky, A. (1990). *Foundations of measurement: Volume III. Representation, axiomatization, and invariance*. New York: Academic Press.

Marcus and Borsboom (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

Mari, L., Wilson, M., and Maul, A. (2021). *Measurement across the sciences*. Springer.

McDonald, R.P (2003). Behavior Domains in Theory and in Practice. Alberta Journal of Educational Research, 49, 212–230.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist 50*: 741–749.

Messick, S. (1998). Consequences of test interpretation and use: the fusion of validity and values in psychological assessment. *ETS Research Report Series 1998* (2): i–32.

Mislevy, R. and Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology (PADI Technical Report 9)*. Menlo Park, CA: SRI International.

Mislevy, R. and Riconscente, M. (2006). Evidence-centered assessment design. In: *Handbook of test development* (ed. S.M. Downing and T.M. Haladyna). Lawrence Erlbaum Associates.

Mislevy, R., Hamel, L. et al. (2003). *Design patterns for assessing science inquiry (PADI Technical Report 1)*. Menlo Park, CA: SRI International.

Riconscente, M.M., Mislevy, R.J., and Corrigan, S. (2016). Evidence-centered design. In: *Handbook of test development*, vol. *vol. II* (ed. S. Lane, M.R. Raymond, and T.M. Haladyna), 40–63. Routledge.

Simon, H.A. 1961(1984). *The sciences of the artificial*. The Massachusetts Institute of Technology.

Vendlinski, TD, Haertel, G, and Mislevy, R., et al+10 (2013) *Using principled assessment design in inquiry (PADI) some frequently asked questions (*PADI Technical Report 12*)*. Menlo Park, CA: SRI International.

Wang, C. (1993). *Sense and nonsense of statistical inference: Controversy, misuse, and subtlety*. Basel: Marcel Dekker.

Zalles, D., Haertel, G., and Mislevy, R. (2010). *Using evidence-centered design to support assessment, design, and validation of learning progressions (PADI Technical Report 10)*. Menlo Park, CA: SRI International

5

Content Validity, Item Response Theory, and Latent Variables

CHAPTER MENU

5.1 Design and Initial Validation of Latent Factors and Domain Models, 192
- 5.1.1 (§49) A Dissertation-based Example of a Model of Progressive Attainment, 192
  - 5.1.1.1 Focus on Internal Aspects of Validity, 193
  - 5.1.1.2 Multiple Revisions Required, 193
  - 5.1.1.3 Isolating the "FORE" Latent Factor from the "Word Knowledge" Latent Factor, 193
  - 5.1.1.4 Other Models of Progressive Attainment, 194

5.2 Content Validity and Item Response Theory, 195
- 5.2.1 (§50) Causation and Content Validity, 195
  - 5.2.1.1 Validating the Content Aspect of Validity, 195
  - 5.2.1.2 Causal Explanations, 196
  - 5.2.1.3 More on Content Aspects of Validity, 197
- 5.2.2 (§51) Content Validity and Content Homogeneity of Indicators/Subtests, 198
  - 5.2.2.1 Content and Construct Validity, 198
  - 5.2.2.2 Indexes of Expert Agreement, 200
  - 5.2.2.3 Homogeneity, 201
  - 5.2.2.4 Content Homogeneity vs. Unidimensionality, 201
  - 5.2.2.5 Linear Factor Analysis of Categorical Items and "Difficulty Factors", 202
- 5.2.3 (§52) Some Aspects of Item Response Theory and Computer-Adaptive Testing, 204
  - 5.2.3.1 Item Characteristic Curves, 204
  - 5.2.3.2 The Wright Map, 204
  - 5.2.3.3 Computer-adaptive Testing: The High-jumping Metaphor, 206
  - 5.2.3.4 Implementation Examples from the High-jumping Metaphor, 207

5.3 Latent Variable Modeling, 207
- 5.3.1 (§53) Generic Score of Homogeneous Indicators: Rough Approximations Without Latent Variables, 208
  - 5.3.1.1 Test Blueprints—The Opposite of Homogeneity, 209
  - 5.3.1.2 Formula Score, 210
  - 5.3.1.3 Item Analysis and Factor Analysis, 211
- 5.3.2 (§54) The Types of Manifest and Latent Variables in Latent Variable Modeling: Values, Scales, and Probability Distributions, 211
  - 5.3.2.1 Objects to be Modeled, 211
  - 5.3.2.2 Level of Quantification, 212
  - 5.3.2.3 Latent Factors, Latent Traits, and Latent Abilities, 214
  - 5.3.2.4 Latent Class Models, 215
  - 5.3.2.5 General Latent Variable Modeling, 216
  - 5.3.2.6 Uncorrelatedness and Probabilistic Local Independence, 217

5.4 Summary, 218
- 5.4.1 Pitfalls to Avoid in So-called *Content Validation*, 219
- 5.4.2 Latent Variables and Factors, 219

5.5 Study Questions, 220

References, 222

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

This chapter includes something about each of the main components of the entire book: psychometrics, validity of measures, test theory, and latent variables. As forecast in Section 3.1 (Chapter 3), we are once again deeply engaged with validity—this time focusing on its content aspects—and will also revisit and expand upon domain theory. This term was introduced to psychometrics by Samuel Messick, who reflects his insights into the trajectory of the study of Content and Construct Validity (see Section 4.1 of Chapter 4). In our demonstration of domain theory, we wish to provide an instance of a study that uses ostensive demonstration[1] with rating scales for a judged performance. We will use our discussion of this dissertation-based domain model as our own demonstration of how models can be created that lead to a domain-specific theory of progressive attainment. It also illustrates the aspects of internal validity—content, structure, and substantive process as explained by Messick (1995). *Ostensive demonstration* is a term Petr introduced to us and is of considerable importance in this book.

We have also added what many call *Modern Test Theory*, or *Item Response Theory (IRT)*. It is introduced briefly in Section 5.1, in conjunction with content validity and computer-adaptive testing. A full treatment of IRT is beyond the ambitions of this book, but we provide an introductory conceptual account of this important subject in Section §52. More importantly, a reader's personal competencies in applying IRT as an aspiring psychometrician will be substantially enhanced by reading and trying out Computational Guide #2 on IRT. You can use either the free version of the powerful and versatile MPlus software package or the full version of the software at your institution. We have amplified Petr's discussion of the topic because IRT has evolved so much and become so central in the time since he wrote his original manuscript.

In addition to Validity and Test Theory (Chapters 3 and 4), two central topics of this book, a third central topic, the *Latent Factor Model*, is addressed again in the second half of this chapter (5.2). This expands on what was first introduced in Chapter 3 (3.3, Section §34) and is a prelude to the full discussion of the latent factor model in Chapter 6. At the end of this chapter, the senior author touches on several additional latent factor topics, including latent traits, latent abilities, and latent class analysis. In this chapter, and in the previous two, we bring the models and expectations of Petr's scientific yearnings farther up to date, as the IRT and latent factor studies are more intermeshed in 2025 than in Petr's day.

5.1 Design and Initial Validation of Latent Factors and Domain Models

5.1.1 (§49) A Dissertation-based Example of a Model of Progressive Attainment

Reo McBride was a doctoral candidate in Instructional Psychology and Technology. It is a design-oriented department previously called Instructional Science. Reo had years of experience in teaching reading to children of all ages, many of whom were not doing well. His father and family had worked for many years using innovative reading methods to help children learn to read more effectively. Reo had personally experienced helping students succeed whom the schools had written off

1 Ostensive definition and demonstration are defined in Chapter 8. It is a method of defining criteria by pointing, or by having an apprentice watch while you perform the intended task. An example of ostensive demonstration is demonstrating the measurement of a segment of fenceline by actually performing successive adjacent placements of a yardstick while being observed.

as "hard to work with" and unlikely to become good readers. Most of them displayed problems in reading out loud, problems with which Reo had become quite familiar. Reading out loud is a familiar example of ostensive demonstration still used in elementary schools. Each student is to stand and read a selection aloud before a small audience with the teacher serving as the judge/coach and with sympathetic students sometimes helping.

5.1.1.1 Focus on Internal Aspects of Validity

Armed with an understanding of validity, and the importance of centering design on obtaining validation evidence for each of the key aspects in Messick's (1995) six aspects of construct validity, Reo sought to design an instrument using graded reading tasks in a manner that would lead to good evidence for validity. He focused on three of Messick's aspects of *internal construct validity* (content, structure, and substantive process). He sought to build a foundation of validation evidence for each of the three, by using evidence obtained through his department-required series of several predissertation studies leading to a final dissertation study. Learning from his classes and advisors about which psychometric tools to use, he set out to develop, implement, and psychometrically evaluate a rating scale instrument based on the latent factors and IRT scaling for a construct he called FORE, Fluent Oral Reading with Expression.

5.1.1.2 Multiple Revisions Required

He documented the constructs he thought should be measured (based on a literature search, his personal experience, and two successful reading programs that he knew well). Next, he went through five iterations of his rating scale instrument, trying it out each time in small studies with teachers and McBride raters, while ensuring that students were reading selections appropriate to their grade level. He revised the scales each time and used the 5th version of the rating scales in his dissertation. Videotaping each student, he selected four raters (who had experience with the constructs) to rate the recordings, with training and adjudication of differences along the way. These raters produced reliable ratings of the constructs across the 200 student protocols. They also performed similarly in the *raters* facet in his IRT study, showing small differences among their ratings of leniency and severity. IRT was used after discovering an oblique rotation of these two factors (FORE and Accuracy) which are correlated ($r = 0.55$), as would be expected since both substantive processes are going on during the task of Oral Reading. IRT was used in addition to the analysis of latent factors, to provide a more powerful scaling solution to order the subconstructs of FORE.

5.1.1.3 Isolating the "FORE" Latent Factor from the "Word Knowledge" Latent Factor

In order to assure that word knowledge would not confound the interpretation of the FORE ratings, his standardized methodology made sure that the reading content was at or below the reader's ability. He had developed a tentative domain theory of progressive attainments with causal implications; that is, he hypothesized that the five main FORE constructs were ordered in difficulty because the earlier were prerequisite to, or enabled, the later ones. His order was *smoothness* < *rate* < *phrase boundaries* < *expression* < *confidence*. If validated, his initial domain theory of progressive attainments provides tentative theoretical explanations based on the substantive nature of each subconstruct. Each one is essential for higher attainment of more difficult ones, while they themselves are enabled by lower ones.

The structure he found met his hypothesis of two latent factor variables and provided evidence of the structural (two-dimensional) aspect of the FORE domain. The order of difficulty, interpreted as progressive attainment, was revealed by IRT scaling. It was *phrase boundaries* < *smoothness* < *rate* < *confidence* < *expression*. This provided explanatory information of underlying substantive process

meanings and enabled teachers as "coaches" to guide individual students by increasing their understanding of the substantive nature of the latent variable of increasing expertise in FORE.

This contributed to a form of domain theory being developed by a group at Brigham Young University and the EduMetrics Institute—a *domain theory of progressive attainment*. Five articles on various aspects of this topic were presented in the 11th annual International Objective Measurement Workshop in New Orleans in 2002. Four of these were eventually published in Garner, Engelhard, Fisher, and Wilson's (2010) *Advances in Rasch Measurement*, Volume 1.

In 2002, when the Objective Measurement Workshop was held, evidence-centered design (ECD) was becoming more and more visible. By the time the McBride dissertation was accepted in 1995, ECD and later the PADI on-line tools were becoming quite visible, and it was apparent that they offered a complete design methodology, a design language that attracted a great number of practitioners, and informational materials that were easy to access. It was not until the research leading to this book, however, that it became clear to these authors that Mislevy, while supporting much of the work of his ETS colleague Messick, was really replacing "construct validity" in every facet of the validity model by developing multiple domain-specific models and using model-based reasoning through continued testing and evaluation. Such domain-specific models are more useful for scientific progress than "so-called constructs," although "construct" is still relevant as a verb because the models are *constructed* by an intentional design process and the scientific philosophy could be referred to as constructivist-realist. "Construct" as a noun is likely to continue to diminish among domain-specific model developers.

Of significance to the research team studying domain theory, the order of progression of the rating scales provided an instructive example of a Domain-specific *Learning Theory of Progressive Attainment*. The original 2002 paper on developing it was in press until the conference book's publication in 2011 (Bunderson, 2011). This paper was concerned with measurement theory and domain-specific learning theories of progressive attainment. A chapter in an earlier book by Bunderson, Wiley, and McBride (2009) emphasized the benefits accruing to instructional design and teaching/coaching individuals from a clear and easily understandable model of progressive attainment.

5.1.1.4 Other Models of Progressive Attainment

The topic of *learning progression* was defined in 2009 by a distinguished panel, as reported by Zalles, Haertel, and Mislevy (2010) in Report 10 of the SRI PADI website. This report applies ECD to designing and developing a scientifically disciplined *learning progression*. A learning progression and a domain theory of progressive attainment are both instances of valuable design language terms that ECD has contributed to the field of assessment: *"... in science learning progressions, the characteristics of what is in ECD referred to as 'Student Model' are the students' evolving reasoning capacities about the focal scientific phenomena." (Zalles, Haertel, & Mislevy,* 2010*, p. 2).*

The example they used to illustrate developing a learning progression was genetics spanning grade bands 5–6, 7–8, and 9–10. They illustrate that assessments developed using their design pattern can separate the linguistic content from the model-based reasoning the learners must master at increasing levels of complexity as the science concepts become more complex. The example does not provide data on students repeatedly performing a similar ostensive performance, but a study cited by Schwarz et al. (2009) does, giving repeated attempts at drawing models and doing model-based reasoning about the evaporation and condensation processes in the water cycle. The repeated performance demonstration of redrawing their original models and explaining them showed evidence of model-based reasoning in individual students. It also showed that teachers

orchestrating this student work can coach individuals after each repetition. Students thus begin to recognize that scientific models can and should be improved.

5.2 Content Validity and Item Response Theory

5.2.1 (§50) Causation and Content Validity

5.2.1.1 Validating the Content Aspect of Validity

We now pick up our discussion of the complete unified model of validity from Chapter 3 and supply the discussion of the Content Aspect of validity postponed until this chapter. Content is often represented as a table of contents. Content analysts seek respected outlines with terms describing the content coverage included within the domain boundaries. This is a linguistic activity, and the type of knowledge—"declarative knowledge"—involves knowing the meanings of the terms and basic concepts descriptive of the way people in the domain write or talk about it. Impressive textbooks used in content analyses may prove quite influential in the final content description in a domain analysis during development. Some of the content categories identified during content analysis will require substantive cognitive processes crucial to criterion performance. These can be described with words and diagrams but performing them is the most important thing to be measured. It is much easier to assess declarative knowledge than performances, and questions are not the best tool for assessing performances. Ratings of observed performances are usually much better. Unfortunately, the usage of the word "content" in psychometric literature often mixes declarative and procedural knowledge together. Expertise (that is, *valuable* expertise) in the target domain usually demands more complex forms of thinking than remembering the meanings of terms or being able to classify additional exemplars under concept labels.

Perhaps over time, it will become best practice to separate the linguistic content from the substantive processing content, as is suggested in Chapter 8 in the definition of a repeatable ostensive demonstration. When Messick lists content along with substantive processes under internal validation, he tells us that substantive domain processes are different from "content." In the usual test-development practices where "content coverage" is central, test "blueprints" often conflate declarative linguistic knowledge (*know-what*) with performance (*know-how*). Tests built from blueprints that mingle these levels of learning often produce a morass of multidimensionality. This multidimensionality is revealed (by factor analysis of the item responses) to provide evidence for latent trait dimensions.

Extensive literature on content aspects of validation exists. We recommend that you learn more than we can cover here. An authoritative resource is Downing and Haladyna's (2006) eEdition *Handbook of Test Development*. In their Book,[2] one can read Kane's chapter 7 *Content-related validity evidence*. As discussed in Chapter 3, Kane is a leading voice in the dominant viewpoint on validity. The validation of interpretations and actions is going to remain fundamental. But also recall our arguments against Messick, Kane, Sireci, and others who do not consider the instrument a part of validity. That limitation was rejected in Section 3.1 of Chapter 3. But they all know how to use

2 This book is out of print. During the writing of this book, the Handbook publishers Taylor and Francis made an eBook available for free download. Now they provide gratis only a preview that includes the first two chapters and unformatted chapter references. Use your college library or interlibrary loan to obtain a copy.

content aspects of validity in a validation argument, and Kane is thorough and complete in his logic, references, and discussion. Consider these initial words:

> Content-related validity evidence is closely tied to test development. The proposed interpretation and use of test scores guides the development of the test and the inferences leading from test scores to conclusions and decisions ... much of the evidence needed to support the interpretations of test scores as measures of observable attributes may be generated during test development as the content domain for the test is specified, data collection procedures are defined, and samples of tasks and conditions of observations are drawn (Kane, 2006, p. 130).

Kane has served on the National Conference of Bar Examiners but is now, at the time of this writing, in one of the five ETS endowed chairs, along with Robert Mislevy. But while he talks extensively about investigating the target domain, he did not cite any of the work done on domain models by Mislevy and his associates, even though they also had a chapter in this 2006 Handbook (Mislevy and Riconscente, 2006).

Kane's chapter 7 gives a brief but very good historical overview of content validation, and a thorough treatment of the concept of validation argument, of which he is perhaps the best expert witness one could call upon to explain where that concept is today. He shows how the content aspect supports validation argument and what are its weaknesses.

In so doing, he supports the logic of using expert judgments to support what content is to be in the target domain. In Section §52 below, Petr – backed by his successor authors – offers cautions about expert judgment as usually practiced.

Like Kane, Petr Blahuš did not quote Mislevy or other works on assessment design but wrote in his own way about the importance of design in his manuscript. Here is one of Petr's comments about latent variables that implies research design in advance: *"Sometimes the lack of a deeper elaborated theory and a well-founded hypothesis causes the concepts modeled by latent variables to be unclear from the very beginning."*

5.2.1.2 Causal Explanations

As Petr Blahuš has stressed in Sections §36 and §37 of Chapter 3, getting to causal explanations is a high goal of the careful scientist in developing and using scientific variables. Borsboom, Mellenbergh, and van Heerden (2004) regard causation as the simple minimal core of validity. Borsboom et al.'s incisive and influential paper was found in Petr's references, and in Petr's own desire to seek causal explanations when possible. McBride and his advisers knew that proving causation is often not feasible because unknown causes may always be lurking in the background, but they saw it as a desirable goal for teachers and coaches to be able to explain the meaning of domain stories of progressive attainment to learners. They recognized, for example, that *looking ahead* led to *smoothness* and a faster *rate*.

Causal explanations anticipated the proposed international metrological definition of measurement in Mari, Wilson, and Maul (2021, p. 37) as summarized in Chapters 3 and 4. A deliberately designed measurement process produces changes that result in an instrument reading. This reading is then converted into a value on the instrument's scale and, through calibration, into a value on a standardized public scale. Causation is found in the process of *transduction*, which in many STEM domains uses physical or chemical sensors that react proportionally to the property of an object that comes in contact with the specifically designed sensors.

We have stressed that human domain-specific scientific models and theories are needed to develop the technology for the appropriate human assessment "sensors." It takes a substantial period of research to produce the items and/or rating scales that work in each specific human domain. The pathway toward developing the instruments that have the information to calculate an individual student's vector of item or rating numbers is through many iterations of sample instruments and tryouts. Samples of thousands of persons are desirable to equip the instrument with stable parameter values that will enable it to provide valid and reliable score values for each person.

The authors of this book have been making a case that principled (theory-guided) design of both instruments and their implementation in both research and applied settings is necessary. Then instruments can deliver evidence whose collection for initial internal validation was designed from the beginning. The evidence argues that the models are meaningful, the scores interpretable, and the actions taken on the basis of the scores defensible.

5.2.1.3 More on Content Aspects of Validity

Looking ahead to Petr's concerns in Section §51, successor authors are in full agreement with his points about the nonscientific nature of using groups of subject matter experts both to judge items and to provide an index of agreement among the experts. This practice seems far from what instrument designers are after: they seek ways to understand what human proficiency constructs are actually operative in test takers when confronted with appropriate domain tasks. In addition to Petr's concerns in Section §51, we add the observation that starting with creating or selecting items before the domain models are understood – before the task- or student-models are even developed – is to not undertand the entire conceptual assessment framework. It is to start in a state of confusion about the substantive processes and dimensional structure of what the domain contains. Such carelessness is likely to miss possible task structures that will best make contact with the oft-invisible property behind the construct.

Mislevy (2009) has cautioned us that invisible traits may exist chiefly in the minds of psychometricians and not necessarily in the minds of the tested. Instead, socio-cognitive theory provides evidence that different individuals with different backgrounds and purposes assemble in their minds information processes at the time of testing. With decent model fit, these appear to developers "as if" there were actually invisible traits already operating within each person as they confront the tasks (see Mislevy, 2018, pp. 15–16, 136–139).

We reply that "as if" is only a starting place. If the latent ability model fit is good, and if we are using instruments offering repeatable ostensive demonstrations, practice with coaching guidance leads to a greater convergence of cognitive processes across persons. The more practice, the better the model fit. If not, since the instrument's designers can also "learn" from validation efforts, a better and perhaps differentiated student model of progression may develop. This is scientific metrology—not with pre-existing "traits," but *a well-fitting model* of what is really happening, interpreted in understandable terms. With repeated trials and coaching, it is expected that learners' cognitive or physical processes would converge more and more toward a coached and practiced test-taker domain model so that the statistical model fit would improve. Ongoing data analysis may uncover significant intransitivities in small groups of learners, and with continuing ongoing evaluation and instrument improvement, a small number of alternate pathways could emerge for different subgroups of learners.

5.2.2 (§51) Content Validity and Content Homogeneity of Indicators/Subtests

In the behavioral sciences, we usually dare not assume that groups of tests or test repetitions might qualify as "parallel forms." There is no definite evidence that a cluster of indicators is associated with the same theoretical concept. How can we be sure that somewhere inside each of the different observable attributes, which we selected into a certain cluster, is hidden something common to all of them? However, there is a certain basis for some *empirical support* for such a claim if the empirically observable operations of those indicators are similar from the point of view of the test-task formulation and instructions. Moreover, the observations must take place under comparable conditions for everyone. The external appearance of reactions of the persons being tested must be the right ones to be sensed by the tasks, or else not counted in the item score. Operations carried out by the test administrator must increase fairness and reduce cheating, etc. In short, standardization is required in the design of both instruments and administrations.

In addition, there is perhaps more important *theoretical support,* which is brought in by content analysis of the *substantive-semantic content* of the indicators. For instance, it may be represented partly by the supposed semantic meaning of a test task to the person taking the test. Another instance of theoretical support, perhaps more important the previous one mentioned, comes from the *consistency of the theoretical network* of the mutual interrelations among the theoretical concepts, as well as between theoretical concepts and empirical indicators, in the given scientific theory relevant to the domain.

5.2.2.1 Content and Construct Validity

In the conception presented in this book, the term "content" is automatically presumed as *semantic* content, hence 'construct' validity is a kind of semantic validity, a validity from the point of view of semantic meaning—say semantic similarity of the content of the indicator and the "construct." *Content analysis*—analysis of semantic content—uses "triangulation" and synthesis of multiple areas of expertise from the above-mentioned different theoretical and empirical aspects. If this process is converging to a joint support of the hypothesis that the indicators in a given cluster primarily indicate the same theoretical concept, we may call the indicators *semantic-content homogeneous*, or *content homogeneous*, or more shortly, *homogeneous*. Then the individual indicators in the content-homogeneous cluster are said to have prevailing *content validity* to their common theoretical concept. And vice versa, if tests have *content validity* to the *same* theoretical concept, then they are called *content-homogeneous*. The usual conception of content validity is based on the idea that diagnostic tools should concentrate on indicators that have a *primary* and *prevailing* content validity to only one theoretical concept and that possible relations to other theoretical concepts should be negligible, if any. Nevertheless, this need not be a rule everywhere, as sometimes the interplay of two abilities in carrying out one task might be important.

Since content validity represents an instance of the *nonquantitatively assessable kinds of validity*, it cannot be expressed by any particular numerical index. At most, the procedure of assessment of a certain content analysis by a group of experts can be carried out by certain methods of qualitative research, by methods of brainstorming, and evaluated by mutual agreement of the experts in terms of different coefficients of inter-rater concordance with a belief that all experts perhaps might not be wrong in the same direction.

The idea among many test developers seems to be that the semantic content of a test task, a statement in an attitude inventory, a question in a questionnaire—generally the content of an indicator—cannot be measured by a quantitative index. In response to this idea, some authors still use *highly problematic terms* such as a "*coefficient*" of content validity. In reality, such coefficients or

indices cannot measure appropriateness of semantic content, but they actually measure *agreement of opinion of different experts* about it. Every time, there is uncertainty about the quality of the experts—somebody had to decide in advance who to invite to the panel of experts. There also must be uncertainty about the objectivity of experts' assessments, their various expectations about what content should be covered in the scale, and their different habits of rigorousness. Further, content validity is considered the above-mentioned degree of belonging or similarity to the *content* of a theoretical concept *as it had been pre-defined*. Hence, the opinion of experts will depend on the definition that is offered to them. Their agreement with this definition is being averaged, not the actual difficulty of the questions in an actual representative sample of students.

Moreover, there are further and deeper scientifically oriented problems. If the experts are adherents of different scientific theoretical paradigms, their agreement/disagreement can sometimes be a consequence of the state of affairs in a given branch of science where two or more paradigms may be competing with each other. Therefore, assessment of content validity cannot be solved by a shortcut study based just on a simple and "ready-to-use" exposition of a list of items presented to some would-like-to-be experts, followed by final computation of an index of their mutual agreement. In addition, an arbitrary choice of a sufficient or minimum limit value of such an index cannot work as a final criterion for decisions about which indicators are to be included in the homogeneous domain related to the theoretical concept, and/or excluded from it. The whole conception of the original research project about content validity must be primarily based on a clear *substantive theoretical paradigm* offering the above-mentioned *theoretical support*.

Sometimes the term "content validity" is substituted by more or less equivalent notions such as "content relevance" and "content coverage" (Messick, 1980, p. 1015, and his many followers call it the content aspect of construct validity). It is not quite clear why the term "relevance" should be substituted for the original term "validity." In this branch of terminology, we also find "item relevance to a 'construct'," which stresses the *primary scientific meaning of content validity* as the belonging of indicators to the domain of a certain theoretical concept (i.e. "'construct' relevance" in the modified terminology). On the other hand, content validity may also be used in a more practically oriented meaning related to test construction. Taking the test battery as a whole, there is also an important aspect of its complete content validity for such a practical purpose. Of course, the motivation for using a test battery for predictive validity or other practical utility (Section §29) has to be based on a deep theoretical background. Nevertheless, it is the practical purpose that determines which *theoretical concepts* and *additional elements* are to be sampled. Then, this specific type of content validity is no longer related to only a certain theoretical concept, but by

> "how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn" (Cronbach, 1971, p. 444), or
>
> "the behaviors demonstrated in testing constitute a *representative sample* of behaviors to be exhibited in a desired *performance domain*" (Standards ..., 1974, p. 28; italics added).

Note the difference between a "performance domain" and a "'construct' domain." Although the more recent Standards (1999) brought in a completely different conception, in most notions and terms it is clear that the quotations represent a version of content validity different from "construct relevance." In their suggestion for a general change in the conception of validity, Lissitz and Samuelsen (2007) actually considered content validity related to a theoretical concept as

unimportant, and they even say that "construct validity as it currently exists has little to offer to test construction" (Lissitz & Samuelsen, 2007, p. 440). For example, in their specific practical orientation to the concept, they consider *content validity to the content of the curriculum* as items' relevance to the purpose of assessment of a curriculum. More recently, we can find still more specific considerations, (cf. Borsboom et al., 2009 in the edited book organized by Lissitz, 2009). However, the scientific importance of content validity as primarily dealing with the *domain related to a theoretical concept* (the above-mentioned "'construct' relevance") is permanently stressed in most current conceptions (cf. Embretson, 2007, 2010).

5.2.2.2 Indexes of Expert Agreement

As regards the rather technical aspects of different numerical indices that are usually used for assessing the agreement of experts, we may find typically two of them:

- *kappa* coefficient by Cohen (1960; cf. also Brennan & Hays, 1992), and
- *CVR—Content Validity Ratio* by Lawshe (1975), or see also the *CVI Index* in Lynn (1986), and Wynd, Schmidt, and Schaefer (2003).

Kappa was originally created to measure agreement in nominal data in contingency tables. Actually, some older measures of association in contingency tables based on χ^2 might also be used, such as the well-known *Contingency Coefficient* $C = \sqrt{\chi^2/(1+\chi^2)}$. In general applications, kappa is usually and frequently used for assessing agreement of raters in different situations, for assessing test-retest reliability of nominal variables, and especially for "inter-rater reliability." It specifically expresses agreement relative to a chance occurrence of agreement ($\kappa = 0$), or down to the lower possible agreement (down to $\kappa = -1$), or up to the highest possible agreement ($\kappa = +1$). Some authors offer auxiliary kappa categorizations, such as Cicchetti (1984) who suggests Poor <0.40, Fair 0.40–0.59, Good 0.60–0.74, and Excellent 0.75–1.00. As was underscored above, it would be an easy but naive and simplistic approach to take some of these values as a definite limit for the final—or even the only—criterion for decisions for which an indicator is designated as "relevant," and for which one is not. Of course, it would be plain nonsense to test a statistical null hypothesis of whether κ is statistically significant (different from exactly 0 in a population of experts).

CVR represents the proportion of relative agreement (from 0 to 1) with respect to the 100% agreement of all experts in the panel, and on this basis, it is criticized for overestimation (kappa is often recommended as more appropriate). In the case when each of the N experts only expresses his/her opinion "yes/no" on whether a given item is relevant or not, then if more than 50% of the experts agree positively on its relevance, the appropriateness of the item is at least slightly positively supported. Originally, Lawshe (1975) used the term "essential"—meaning that an item or subtest is either an "essential" empirical indicator of the theoretical concept and/or it is "essential" to another purpose. In Lawshe's formula, the absolute frequency of panelists who jointly agree that the given item is "essential" is symbolized by n_e—thus the 50% limit might be simply assessed by the proportion n_e/N. However, Lawshe (1975) recommended another expression: $CVR = (n_e - N/2)/(N/2)$. This content validity ratio varies from +1 with 100% agreement, through 0 for 50% agreement to –1 if there is no agreement and $n_e = 0$. Since *CVR* is perfectly derivable from the proportion n_e/N, some motivation for using this complicated version might be found in its similarity to possible values of the correlation coefficient or Cohen's kappa.

Typically, the raters' opinion about "relevance" of each of the indicators is expressed on a Likert scale, i.e. on an instance of an ordinal scale. Therefore, it is not quite clear why so often kappa is used as it was created especially for nominal categories. Hence, it does not reflect information about

agreement of order. From this point of view, the well-known *Kendall's coefficient of concordance W* (e.g. Siegel & Castellan, 1988) would be more appropriate, but probably due to some tradition, it is not frequently used in the assessments of content validity.

5.2.2.3 Homogeneity

In the framework presented here, the term "homogeneity" is always considered content homogeneity, i.e. in the sense of the semantic content of behavior. This notion must be carefully *distinguished* from other possible connotations like:

- *homogeneity of correlations—a concept close to the unclear notion of consistency,*
- *homogeneity of an item with the total score,*
- *homogeneity of a binary item's correlation with respect to its maximum,*
- *homogeneity of items with respect to their difficulty ordering* (related to Guttman's Ideal Scale),
- *homogeneity of items with respect to their nonparametric latent trait* (Mokken's analog)
- *homogeneity as Loevinger's scalability* (a kind of proximity to one latent trait)
- *homogeneity of errors,*
- *homogeneity of variances,* known as *homoscedasticity,* i.e. equality of variances, and finally
- *factorially homogeneous tests,* referring to a test that exhibits unidimensionality according to the latent common factor model (Spearmanian *g* factor).

All the above-indicated meanings can be found in the literature that deals with measurement in behavioral and social sciences, psychometrics, and statistics. Especially, the last term, which deals with the syntactic-formal meaning of mathematical-geometrical unidimensionality, should be distinguished from the semantic-substantive meaning of content validity. This problem is now discussed in the next subsection.

5.2.2.4 Content Homogeneity vs. Unidimensionality

In this text, "unidimensionality" of indicators (subtests) is considered only as a term in the language of mathematics and statistics. This term is especially motivated by the geometric interpretation of the latent variables in most mathematical models. However, some authors conceptualize content homogeneity and unidimensionality as synonymous, while others separate the two implicitly or explicitly. For example, briefly interpreting the paper by Mulaik and McDonald (1978), Mulaik (2005, p. 195) mentions the problem that in

> "selecting homogeneous sets of variables by factor analysis .... [t]he moral is that *factor analysis alone should not be the criterion of homogeneity* of items. There must be *external criteria of content* ... to guide one to retain or exclude variables from considerations as *homogeneous*." [Italics added.]

We insist that it is necessary to *distinguish the modeled object*, here the substantive/semantic content, from *its model*, here the mathematical latent variable as its syntactic/formal counterpart. Compare Table 3.3 and Section §39 which attempt to give order to the mishmash of terminology in this area. In this connection, other mathematical models for latent variables besides linear factors have gained widespread use since Lord and Novick's (1968) influential volume on *Statistical Theories of Mental Test Scores*. This book, along with many articles preceding and following it, has shown how item characteristic functions can be estimated and graphed. Examples were given in Section 5.3 and in our Computational Guide #3. Item and test information curves are also of wide use in assembling and investigating tests. In modeling latent traits with nonlinear

item and person characteristic functions, both normal ogive and logistic ogive nonlinear models are defined that relate the probability of correct or incorrect item responses to an underlying latent trait.

Before IRT models came into wider use, exploratory linear factor analysis had been applied to binary items yielding one ample general factor that was still not sufficient to fully "explain" the inter-item (phi) correlations. Sometimes spurious "difficulty factors" were found at the lower and upper ends of the scale, where criteria like eigenvalues of 1.0 and greater led to the possibility of more than one latent trait factor. In cases where the construct is well understood and the indicators represent it (e.g. as in the vignette at the beginning of this chapter), the low- or high-difficulty clusters were conceptually homogeneous with a well-specified and well-understood progression of higher to lower abilities. With strong theoretical rationales in place, the latent trait will be seen to be unidimensional. IRT is both a complement to and a continuation of classical test theory (CTT).

We will deal with the terminological issues later (including the so-called "consistency" of indicators) but let me outline its importance with some examples right now. One problem is that a heap of variables grabbed without any joint content of meaning might be formally unidimensional in the sense of having the correlations "explained" by just one latent common factor almost perfectly. It could happen purely by chance and/or also in the case without any prior content hypothesis as in exploratory factor analysis (EFA), when the "discovered" factor has no a priori semantic interpretation, and/or an interpretation is brought in post hoc by willful speculation. Alternatively, it may occur due to quite different semantic content, other than that formerly intended when the indicators could very well measure a different "construct," say a quite different aspect or facet of behavior.

Furthermore, formal unidimensionality may be attributable to an influence that is better studied by another branch of science operating at a different level of resolution. For example, unidimensionality arising from macro-behavioral aspects is fundamentally different from that resulting from neurophysiological or micro-biochemical aspects. In addition, unidimensionality can be produced due to a prevailing influence of environmental, situational, or contextual circumstances, or as a consequence of certain diagnostic artifacts that were common to the cluster of indicators (being rather a consequence of a specific character of the diagnostic methods used).

5.2.2.5 Linear Factor Analysis of Categorical Items and "Difficulty Factors"

Unidimensionality can also be related to the *linearity* of a model. One problem connected to this has historically been known as the problem of "difficulty factors." Before IRT models for binary items came into wider use, especially due to F. M. Lord's normal ogive model in 1950 (Lord, 1950), exploratory linear factor analysis had been applied to binary items yielding one ample general factor that was still not sufficient to fully "explain" the inter-item (phi) correlations. Several further factors were needed that often corresponded to items clustered according to different levels of their difficulty, but still with similar semantic content. The number of these factors, sometimes called "spurious due to difficulty," showed higher mathematical dimensionality of the indicators than would have been expected by their semantic-content "genuine" factors. Thus, formally, the items were found to be multidimensional. But, the use of nonlinear, say normal ogive or logistic regressions of the binary items on the latent variable in an IRT model, changed the situation, and were able to represent the items as unidimensional—however with a different, nonlinear meaning. The general factor can then better explain the inter-item (tetrachoric) correlations without distraction due to spurious difficulty factors.

In addition, in the nonlinear polynomial factor models elaborated by R. P. McDonald in the 1960's, the theoretical concept may exhibit its influence on the indicators in several ways simultaneously: On one hand, the "construct" may manifest its influence on the performance score in a test through its linear effect. On the other hand, it may also function through quadratic or cubic effects, which may all be combined, while the substantive semantically homogeneous content of the "construct" stays the same. This type of nonlinearity deals with nonlinear functions mutually relating the latent variables (technically speaking, the model is not nonlinear in its parameters, but is nonlinear only in the variables). These nonlinear models, then, substantially change the traditional way of understanding unidimensionality. For instance, if a linear model works well with two latent common factors, a nonlinear model may do the same with just one latent factor whose parabolic (quadratic function) "version" was substituted for the former second linear factor. Speaking metaphorically, we would need two kinds of "unidimensionality," linear and nonlinear—since in the latter case, we have a kind of a twofold interpretation of what a "*general*" factor may mean.

Another argument for distinguishing between dimensionality and homogeneity is that otherwise it would be necessary to distinguish two variants of the term "homogeneity." The first one reflects a semantic notion—the homogeneity of the substantive theoretic content, i.e. *homogeneity that does not depend on any type of a formal model*. Second, we might consider a kind of "narrower homogeneity," which would be limited just to those instances of semantic homogeneity that could be modeled by formal mathematical unidimensionality in terms of linear LCF models. Some authors assume only this narrower meaning which consequently leads to a conception that homogeneity and unidimensionality were the same—as if the numeric values of the artificial latent variable were actually those levels of a theoretical attribute that is to be modeled (by that mathematical variable). Then, for example, the number constructed from the scores of a battery of reasoning tests by an LV model would be that "very amount of reasoning capacity" hidden in the brain and not just its formal representations in terms of a model. Thus, keeping the *distinction* between the object being *modeled* vs. its *model*, we will distinguish between substantive-semantic homogeneity and its representation by mathematical unidimensionality.

From one point of view, a cluster of indicators, which was found to be content homogeneous, need not be unidimensional in terms of its mathematical model if the cluster needs to be "explained" by more than just one modeling dimension, cf. the case of "difficulty factors." From the other point of view, if a cluster of indicators had been chosen by an a priori hypothesis of semantic-content homogeneity, and its formal latent variable model were laden on an elaborated theory—a well-founded substantive theory consisting of possible linear or nonlinear relationships—then this fact may be recognized as empirical support for the hypothesis. Specifically, if a unidimensional mathematical model formally fits the indicators, it may suggest that the theoretical concept might also be content homogeneous—but not the contrary. In this context, we should also be aware of the distinction between the situations when it is meaningful to apply the notions "correct"/"incorrect" and when it is not. Given that a certain unidimensional latent variable model has been fitted to the data, unidimensionality should be understood in the sense of this very model, i.e. only *in terms of the given model*. Hence, it is not questionable and cannot be "incorrect" (perhaps just up to some computation errors). On the other hand, the sentence "these (unidimensional) indicators are content homogeneous," when assessed as an already given statement, may still be meaningfully judged as correct and/or incorrect in the light of its semantic interpretation.

Another aspect, touched here only in a preliminary way, deals with the problem of the *dimensionality in hierarchical models* of rather complex scientific theories with "constructs" of different order in several layers of hierarchy. Most behavioral theories contain theoretical concepts that are

hierarchically organized in several layers or strata. A classical illustration may be introduced by the still-ceasing sound of Spearman's General Intelligence. Some psychologists still recognize it as an umbrella concept in the highest layer above all other possibly complicated, subordinated structures of intellectual and cognitive abilities—cf. Vernon 1971. Hence, one may think about the possibility of recognizing that different layers of the hierarchy might have different dimensionality, i.e. that *dimensionality is specific to a hierarchy layer.*

The development of Confirmatory Factor Analysis and analytic Target Rotations in factor modeling, as discussed in Chapter 7, are now the preferred methods for testing advanced hypotheses about factorial structure. Without an a priori hypothesis, too little research and thinking has been done, with the results Petr has outlined above. Confirmatory factor analysis is a preferred tool for confirming the number of latent factors, the relationships of the factors to their indicators and to each other, and their connection with the property to be measured. With binary and ordered categorical indicators, IRT (including multidimensional IRT) is now usually the preferred method for generating the scales and assigning the values of the latent factors to individual test takers.

5.2.3 (§52) Some Aspects of Item Response Theory and Computer-Adaptive Testing

Chapter 4 and Computational Guide #1 have introduced and outlined basic principles of CTT. Subsequent developments in educational and psychological measurement have led to a broader and more general modern test theory. The most well-known and widely applied specialty within contemporary test theory is IRT. IRT extends the latent variable or latent trait perspective developed in this book to the explicit modeling of item response probabilities, beyond just analyzing the mean and covariance structure of test items. IRT represents a substantial departure from CTT by improving model-based measurement and making inferences to the latent trait more disciplined. It provides new and more powerful statistical tools and provides proof that some of its models meet the axioms of additive conjoint measurement so that using the interval properties of the real number system can be more fully justified. Computational Guide #2 gives an introductory applied treatment of IRT, providing the basic mathematical representation of key IRT models and their implementation using statistical modeling software with example data. It also presents associated output listings and various possible tables and graphic visualizations. We introduce here some selected concepts and applications of IRT, encouraging readers to pursue the topic more deeply according to their own interests.

5.2.3.1 Item Characteristic Curves

Two useful graphical representations are the logistic item function plot and the Wright map, or item-person map. The logistic function looks like a cumulative normal ogive—a long flat *S* curve (see Figure 5.1). These are commonly referred to as item characteristic curves (ICCs). The computational guide explains the normal ogive as well as the logistic function of ICCs. For a dichotomous or binary item, the item difficulty threshold parameter is the point along the underlying latent trait continuum where the probability of successfully responding to the item reaches 0.50. The discrimination parameter is the slope of the *S curve* (rise over run) at this inflection point.

5.2.3.2 The Wright Map

A Wright Map shows the distribution of Theta (θ) for each person's trait or proficiency estimate opposite the associated values of the difficulty parameters (β_j) for each item. These two parameters are compared on the same vertical log interval (logit) scale to give the analyst or later reader a perspective on the relationship between them (see Figure 5.2). The rotated histogram of the estimated

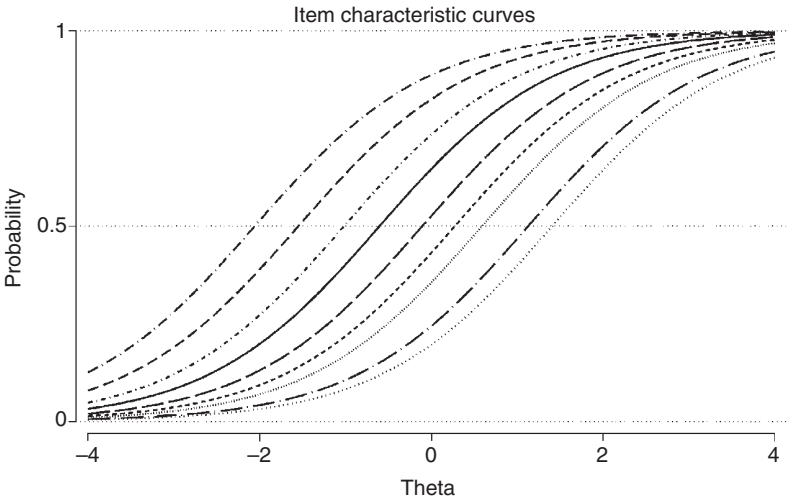


Figure 5.1 Item Characteristic Curves (ICCs) for nine binary test items based on the Rasch model from *Item Response Theory* (IRT).

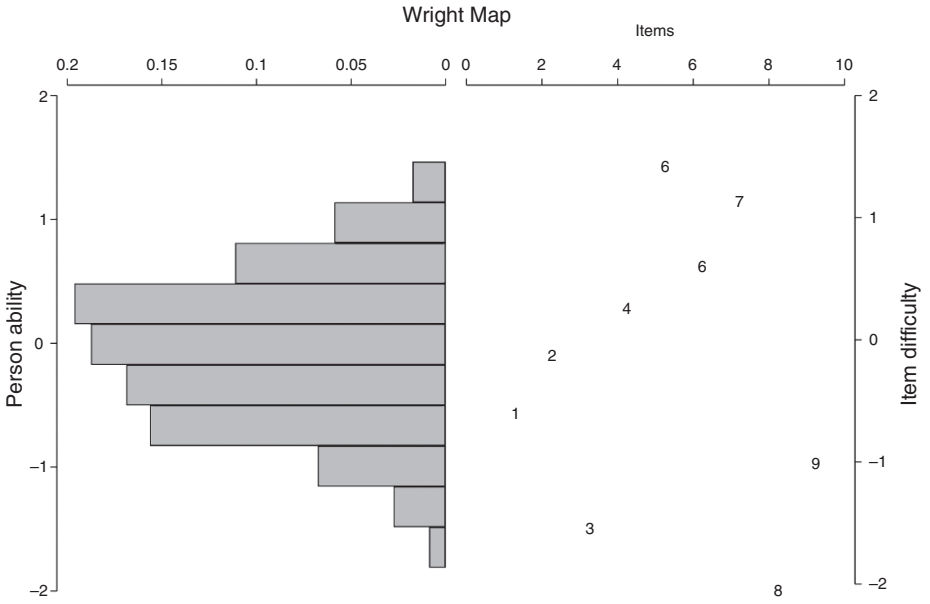


Figure 5.2 A Wright map for nine binary test items.

Theta values is shown in the left panel of the figure while the plotted item difficulty parameters are shown in the right panel. These relationships hold for well-fitting unidimensional Rasch models. The representation called a Wright Map was introduced by Wright and Master (1981). It was referred to as Wright Map by Benjamin Wright's doctoral student, Mark Wilson. It honors Wright and is much easier to talk about than a *vertical person x item parameter plot*. It is a familiar term in Rasch model-using psychometric circles. Benjamin Wright sought out and studied with Georg

Rasch and did much to popularize the Rasch model. He influenced many doctoral students and degreed professionals who became effective users and proponents. It has been proved mathematically that if a set of data meets the conditions of the Rasch model, it fits the axioms of additive conjoint measurement and thus supplies quantitative scales at the interval level (Newby, Connor, Grant, & Bunderson, 2010).

The Wright Map in Figure 5.2 plots parameters for the same nine binary items for which the ICCs were displayed in Figure 5.1. Notice that items 1, 2, and 4 have difficulty values near the center of the Theta distribution, while the remaining items have higher or lower difficulties. The Wright Map facilitates evaluating or selecting items with respect to their positioning along the Theta continuum.

Variations on the Wright Map can also be constructed for well-fitting unidimensional polytomous Rasch models (i.e. the Rasch version of the Partial Credit Model). In this case, multiple item threshold parameters are plotted for each item rather than the single difficulty parameter for binary items.

5.2.3.3 Computer-adaptive Testing: The High-jumping Metaphor

In the remainder of this section, we use the example of high jumping to illustrate some of the concepts that IRT can help us clarify. One of the common applications of IRT is computer-adaptive testing, an increasingly important application of psychometrics and another reason to pursue the study of IRT. To understand computer-adaptive testing, consider the metaphor of an instrument to assess the skill of a high jumper. This is a metaphor after the heart of our kinanthropometrist senior author, Petr Blahuš, whose many studies in sports and skilled human movements gave him a clear insight into the difference between vague constructs and those with causal explanations.

The instrument for measuring high jumping has two calibrated poles rising to a height greater than what the best high jumpers have reached. A light crossbar rod is placed on the calibrated projections on the back of the apparatus so that a person who did not jump high enough to clear it will knock it off before tumbling into the soft-landing material on the other side of the measurement instrument. There is a calibrated running path, then a mark for take-off. The jumper gets three tries and his or her score is the highest rod placement cleared under these standardized and well-calibrated circumstances. A referee or judge observes to catch all possible faults, including that the plane below the bar is not broken before soaring over the bar, and any contact that knocks the bar off is a failure. The apparatus is a metaphor for the task and person measures appearing on the same scale, like the Wright Map. On the sidebars, one can read the difficulty of the task—the height of the bar. The adaptive nature is that the athlete (during practice at least) gets to choose the difficulty of his/her first attempt. Succeeding at that and still having jumps left, the athlete can have difficulty moving up one or more notches. In this way, the adaptive instrument "adapts" to the ability (Theta) of the person and does not give a lot of tasks that are too easy or too hard tasks.

Obviously, for this to be a useful metaphor for computer-adaptive testing, we must have a replacement for the millimeter scale by which both persons and tasks (items) are scored. To use IRT for this purpose, we must have a way of estimating the parameters of all tasks presented to all candidates. Alternative parameter estimation methods are available in IRT (e.g. Baker & Kim, 2004). They are both alternative statistical approaches and feature purposes that differ. These methods are computationally intensive as they iterate until some measure of good fit is obtained; thus, IRT requires computational power and technical expertise. Access to the best programs has been revolutionized over the 15 or so years since Petr used programs that now are not easily available. Happily, the popular programs M Plus and Stata, which are used in your computational guides, are available at most universities or from vendors, sometimes offering free but limited versions.

Back to the high-jumping metaphor, and many other sports that score performances based on time, distance or height, and force or strength, we have extensive attributes measured by metric or other standardized units. Athletic events which use the SI units for their public scale are modern examples of psychophysics. The development of IRT has depended on computer programs that estimate the best-fitting parameters of the logistic function or normal function (or some others), yielding placement of both person proficiency and task difficulty on the same scale where a unit is one logit (one standard deviation on the logit scale). These logit IRT scores can be converted to other linear scales where standardized conventions allow a zero and a unit other than the logit to be used, so that the scale may be calibrated to a publicly accepted scale.

5.2.3.4 Implementation Examples from the High-jumping Metaphor

In Chapters 3 and 4, the need for *instrument design*, *scaling*, and *implementation design* was explained. For any testing session where the instrument is implemented and used, much planning must be done, and featured in the design of the instrument and its administration. In high jumping, the locations for crossbar placement pre-calibrated on the vertical poles is one of the most straightforward aspects of implementation. But that is not enough. The entire starting position, runway, and landing pit require standardized treatment so that records can be compared from one location to another. Thus, the entire jumping activity is standardized in the take-off distance and in as many ways as feasible to assure comparability of achievements. Any and all construct-irrelevant variance must be removed to prevent clouding the interpretation of the highest rod cleared. We would want to rule out doping, or interference with a jumper during any time prior to or during his ready-run-jump cycle that would degrade an athlete's performance.

Now we can begin to see some of the great differences between classical tests and computer-adaptive tests, using IRT. A classical test requires everyone to perform the same series of tasks—the same number—in the same order. Few jumpers could keep doing their best for 50 items, and it would be slow, inefficient, and wasteful, and fatigue would introduce construct-irrelevant variance. So, high jumping is adaptive to the skill of the one being measured. In practice, the athlete can set the starting level as above. In tournaments, the judge knows who is in the competition and can set one starting height for all. An athlete has the right to decline to attempt at this level and may resign from a competition, shortening the length of the match. Jumpers get three attempts and, if successful on the first or second attempt, can ask the judge to set the bar at a specified higher level on their subsequent one or two remaining attempts. Unlike mental tests where the construct cannot be made visible, the height of the bar is visible to the referee, who reports to the audience. The physical contact with a bar setting higher than the ability of the jumper (and the rules of scoring) causes the referee to assign a score that is the highest bar placement attained prior to any last failure to clear the bar.

5.3 Latent Variable Modeling

Section §53 introduces using construct-appropriate indicators to measure a designated common theoretical concept via latent variable modeling. Since the latent variable is measured only indirectly, a related problem is how to construct estimates of the unobserved latent common factor for each individual. Formula scores combine information from the indicators to roughly approximate the latent variable. This is sometimes done with methods such as getting the sum of the indicator scores, or by summing or averaging the equally weighted or differentially weighted

standardized scores (z-scores). In the special case of binary or dichotomous indicators coded either 0 or 1, the total or number correct score is often used as an estimate of the underlying latent trait. This will usually be an adequate estimate only if the data fit a Rasch IRT model. With the older item analysis method, correlations between the indicators and the number correct score (or total test score) are sometimes used as weights for the indicators to create a weighted composite score. Alternatively, when using factor analysis, the factor loadings or the reciprocals for the indicator residual variances are sometimes used as weights when creating composite scores.

In Section §54, some guidance is provided on how to select an appropriate model. It discusses values, scales, and probability distributions, as well as various levels of quantification from nominal to ordinal and interval. The concepts of latent factors, latent traits, abilities, and latent class models are also covered. Additionally, the crucial issue of uncorrelatedness and the requirement for local independence are discussed.

5.3.1 (§53) Generic Score of Homogeneous Indicators: Rough Approximations Without Latent Variables

Content homogeneous indicators are assumed to indicate the same common theoretical concept. So, in a rather broad sense, all of them are "of the same kind," of the same *genus*. For instance, we might like to construct an agglomerate score of a homogeneous[3] battery of tests, which would help us estimate the level of the theoretical concept better than using just one test. This is somewhat similar to the situation of using replications to make an estimate $\hat{\tau}$ of the specific true score τ. Although the various homogeneous subtests work as indicators having content validity to the same theoretical concept, they cannot be recognized as *replicated* measurements of abstract hypothetical levels. Hence, a problem arises here, namely *how to construct estimates of the hypothetical level* of "ability" in a concrete individual from the knowledge of his/her set of scores on a cluster of indicators. However here, instead of the desired genuine estimation of $\hat{f}$,[4] we often encounter only a somehow constructed particular, possibly quite rough, approximation of the latent variable *f*. This usually happens due to a simplistic, not scientific, approach owing to a long history of using a traditional test battery or inventory. Therefore, it is necessary to employ different symbols and to distinguish such a mere *approximation* $\dot{f}$ of the true *generic score f* from its *genuine estimation* $\hat{f}$ in its statistical, psychometric, and methodological meaning.

We pause at this point to comment on the limits of variously used *content homogeneity* concepts. It was Jane Loevinger (1947, 1948) who made the concept of test homogeneity take root. She observed sensibly that the collection of items in a test should all be measuring the same thing, whether a trait, ability, or processing skill. Petr and his good friend, Roderick McDonald, were fond of this term. By contrast, we successor authors have been talking about something domain-specific and thus much less vague. In a strong domain-specific person or student model designed using principled design methods such as ECD, the indicators do not seem to be homogeneous in any simple sense. They could be measuring different levels of attainment in a progressively evolving ability or skill. Whether there are invisible mental processes or physical processes visible with high-speed photography, these different levels of competence often appear quite distinct, and coaches can

3 Another term besides "homogeneous" that is often used to refer to a set of tests that assess the same construct in a latent factor model is "univocal"—metaphorically, all the indicators are singing the same melody.

4 Note that after talking about a hypothetical level of ability Petr is using *f* for the latent factor while in IRT as mentioned, θ is used for the measureand, usually a person *i*, and β_j is used for the task *j* difficulty parameter. "*f*" is about the factor, usually clusters of items or ratings. Factor scores for persons may be calculated.

explain them separately. They are not homogeneous in "content" but are progressively appearing aspects of a unitary skilled performance.

In the FORE study, for example, the *phrase boundaries*, the *smoothness* (using "look-ahead"), *rate*, *confidence*, and *expression* are distinctly different rating scales and can receive distinct mentoring advice after the measurement event. However, they are all integrated by the common ostensive demonstration into a practiced, smooth, unitary performance, especially at higher criterion levels. Despite the content inhomogeneity, these components of attainment did show up as a single latent factor when the other factor of word knowledge was separated. As Petr has discussed, sometimes, the differences in the component parts of the unitary performance show up as difficulty factors. In earlier try-outs, some of the FORE constructs did so, leading to revisions in the rating scales before the final study. Models of progressive attainment for the persons may seem "bumpy" and not homogeneous, yet they fit together in a growth model. Such models have grown in quality and utility substantially since Petr's interest in the word "homogeneity" was popular. Look at the literature on learning progressions that we have cited and the references within them. We advocate domain-specific model-based reasoning as superior to trying to understand how to classify tasks by their "homogeneity." His examples show situations where this language has been useful in the past. Before returning to our discussion of indirect measurement in latent variable modeling, let us consider another problem area, test blueprints.

5.3.1.1 Test Blueprints—The Opposite of Homogeneity

Expert judgment is not the only "common law practice" to raise our warning flags as defenders of the science of measurement. Another is the test blueprint. Consider the pressures on school educators to "cover" state-mandated content. Postsecondary education has mandates from professional societies. This leads to the practice of constructing test blueprints to make sure the mandated content topics and objectives are all taught. Content coverage plans adhere to these mandates, and tests— or evaluations, as medical educators call them—and rightly so, as they more often include performance evaluations and provide validation evidence for effective teaching in line with the mandates. They can also be used to make sure teaching plans do not leave out mandated content. But test blueprints, while helping manage checking off content coverage, force multiple potential latent factors together and conflate simple knowledge of terms and concepts with higher cognitive processes or with trained physical skills (like drawing blood or inserting and servicing an IV).

A 2009 reference on test blueprinting published in *Medical Teacher* (Coderre, Woloschuk, & McLaughlin, 2009) explains the mandates for a form of content validation, and the uses of test blueprints both to provide validation and educational value.

> "Content validity is a requirement of every evaluation and is achieved when the evaluation content is congruent with the learning objectives and the learning experiences."
>
> "In addition to validating evaluation content, a blueprint can also be used to guide selection of curricular content and learning experiences."

The point we are making is that these well-established practices are mandated, and practitioners try to cover their mandates (yet practitioners get educational value out of them through blueprints). Measurement as discussed and defined in this book is not a tool for showing conformance to mandated content across mixed domains and mixed competencies. These blueprints, representations of the content, cannot function as measures of scientific latent factor dimensions. As such, they conflate many dimensions together, most of them dimly understood at best. This competes with the search for scientific regularities in the growth process of these learners that might lead to better effectiveness and progress for each of them. It flies in the face of the need for appropriate

indicators to measure a designated common theoretical concept via latent modeling. The use of blueprints in the planning of tests/evaluations is not metrology but management. This deliberate conflation of two divergent purposes is described by Raymond and Grande (2019), also in the *Medical Teacher* journal.

"The steps include identifying the major knowledge and skill domains (i.e. competencies); ... and establishing the amount of emphasis to allocate to each knowledge or skill domain."

Like a cooking recipe, the blueprint specifies in advance a mixture of knowledge and skill domains and how much of each ingredient is to be mixed in. We now return to the topic of latent variable modeling.

5.3.1.2 Formula Score

There are several rough methods, usually based on an unclear intuition or convention in the form of written "instructions for preparing instant soup"—they are usually very simple. They may be part of a test manual representing an intuitive or habitual practice in a certain area of using the test or questionnaire. These procedures usually consist of a set of not clearly justified numerical computations. Such a computation may actually be applied to any bunch of observed scores, often without testing their suitability—one just has to believe in it. Let's call these methods of constructing a generic *composite score* by definition, a *"formula score."*[5]

One type of formula score is the *composite score* z_c or z_{cw}. This is a simple and intuitive method for constructing a rough approximation of the generic score. It could follow the prototype of a *composite score*, the simple sum, or a weighted average (as described in Sections §23 and §35). When using a *weighted generic composite score* z_{cw}, the weights w_j can be based on different types of rational (or irrational). Sometimes the weights w_j can be based on the different reliabilities of the separate subtests. For that purpose, the weights could be chosen as reciprocals of the separate *error variances* of each of the subtests (which would follow the concept of Fisher's information I). As we will see later, even such weights are not optimal. Moreover, of course, the composite simple sum-score z_c of four verbal tests, for example, would be only a very crude approximation of the "level of verbal ability" comparing it to the rationally based modeling of f in terms of latent-variable models.

Another kind of formula score is the *item raw-score sum*. The terminology is somewhat ambiguous in this area. I prefer the terminology using *subtests* and *test batteries* since I understand a *test* as a general name for *any standardized empirical research/diagnostic method treated by test theory*. In the usual framework, an "item" is just a special version of a subtest, even a dichotomous one (i.e., whose two alternative scores are 0 or 1). On the other hand, there is also a widely used habit that a *battery of items* is also called a *test* whose *test score* is the simple sum of the item scores, i.e. the number of 1's with binary items, say the number of correctly answered questions. It still leaves an open question: To what extent could this raw score be recognized as a *good* approximation of the level f of a theoretical concept? The problem is similar to the above-mentioned simple sum or simple average of z-standardized subtests that are not dichotomous.

---

5 The notion of "*formula test score*" was discussed by R. P. McDonald in about 1999 for a kind of *raw-score of homogeneous items*, typically a weighted sum of the zero/one items. It could also be a quite sophisticated function, including an estimation of a latent variable in IRT models. Actually, any genuine and sophisticated *estimation* of latent variable scores, including the latent common factor score, is computed by a formula that is mathematically rigorous and conceptually sound. Bartlett's method for maximum likelihood (ML) estimates of latent factor scores f is also computed by a certain "practical computational formula." (See §58 in Chapter 6 for more information on the maximum likelihood method.)

5.3.1.3 Item Analysis and Factor Analysis

Another approach is the so-called *item analysis*. It belongs to the family of methods that are slightly more complicated technically, a fact which might obscure that they are not based on a deeper methodological conception. One of its typical versions of "directions for use" is: Compute the composite simple-sum score, the so-called test total. Then, compute simple correlations between each of the items and the composite as an internal validity of the items to the test total. Finally, multiply the item scores by these correlations as the weights w_j and compute a new composite—the weighted-sum score. When the weights are all equal to 1, of course, the weighted composite becomes the simple total score.

There are also mathematically sophisticated types of factor scores, many of which are embodied in commercial statistical packages for "*factor analysis*." This could be in any one of the several forms of EFA which are presented under this ambiguous and confusing name, especially in the forms of *principal factor* analysis and/or *principal components analysis* (PCA). However, there are some other methods in the fuzzy family under the umbrella name of "factor analysis" which suffer from similar problems. Five types of such problems are:

1) They do not represent falsifiable hypothesized models; they are nontestable tautologies whose numerical computations may be applied to any correlation matrix (actually, to any symmetric matrix).
2) They are not primarily focused on "explaining" *correlations* under the scope of the principle of local uncorrelatedness; rather, they "extract" factors/components that at the same time "explain" a certain percentage of the *variance*, leaving some arbitrary proportion as "unimportant," including some "unimportant" part of the correlations.
3) Therefore, the methods need as many principal factors/components as there were observed original variables to "explain" the correlations completely, even in the case where a genuine latent variable model (such as the LCF model) can offer a complete "explanation" even by only a single latent common factor.
4) They carry out a nonstatistical (deterministic, algebraic) transformation, usually orthogonalization, of the original observable variables into "principal components/factors," which, moreover, are not genuinely latent in the sense of the latent common factor model.
5) They do not allow a confirmatory approach to verification or falsification of a preliminary hypothetical structure.

5.3.2 (§54) The Types of Manifest and Latent Variables in Latent Variable Modeling: Values, Scales, and Probability Distributions

There are several main aspects at our disposal for a conceptually based choice of a latent variable model to be used, which depend mostly on the character of the manifest as well as the latent variables, their level of measurement, the form of the probability distribution, and on the type of association functions.

5.3.2.1 Objects to be Modeled

The first possible major consideration in choosing a modeling strategy is *the type of "objects" to be modeled*, that is, the observables that are to be reproduced by the latent model, which actually may be either:

x—the observed *scores themselves*, i.e. *values* of a manifest variable, in the discrete case, its different individual values x_k whose possible number is m (the sample space of x)

and/or:

$p(x)$ the *probabilities of the scores* approximated by the *proportions* $p(x_k)$ of subjects with the same x_k; it means that the probability $p(x = x_k^*)$ that a randomly chosen subject would have his/her score equal to a certain given value x_k^* is approximated by a *relative frequency* $p(x_k^*)$.

Up to now, we have not mentioned the second type to keep the presentation narrow and simple. The models with latent variables of that second type are not focused on reproducing the original scores x in the observed measurement units—inches, seconds, pounds, or score points—but they are aimed at modeling their observed frequencies $p(x)$. Typically, this is the case when the indicators are purely qualitative nominal research variables or dichotomously categorized ordinal research variables whose values are coded in terms of quasi-quantification by (0; 1). Formally, the traditional dichotomous item is just a special case of that way of scoring (i.e., "1" for "Passed the test", "0" for "Failed").

There is an interesting relationship between treating the two alternatives—either modeling the original scores x and/or their $p(x)$ as two different modeled "objects." It consists of some kinds of equivalence of mathematical treatment as well as of interpretation due to the quasi-quantification by (0; 1). We see that item difficulty now becomes $p_x = p(x) = \overline{x} = E(x)$. Again, because of the (0; 1) quasi-quantification, an analogy also holds for the conditional mean (that is, the conditional expectation), and the conditional proportions that stand for the approximated conditional probabilities for a given latent level f yield $p(x \mid f) = \overline{x} \mid f = E(x = 1 \mid f)$. The last symbol on the right is an instance of *regression* by its very definition, and the conditional mean score of subjects with the same "located" level of "latent ability" f is then the same as the proportion of these subjects with the same level f who passed the item x successfully (coded by $x = 1$). This proportion may be supposed to approximate the conditional probability $p(x = 1|f)$ that a person taking the test having ability level f would be successful in passing the item x.

Since most latent models are based on regression functions, the above-mentioned mutual relationships allow one to generalize them and formulate a very general latent variable model that includes quantitative as well as qualitative variables, both manifest and latent—the *generalized latent variable model* (McDonald 1962). Although the mathematical conception is formally equivalent for this family of models, the specific kinds of motivation for their creation in the past were different as well as their present practical applications.

5.3.2.2 Level of Quantification

The second major consideration in choosing a latent-variable modeling strategy is *level of quantification*. The level of quantification of the latent variables and their type of *measurement scales* reflects their character, which depends on the level of quantification of the whole corresponding theory and its associated theoretical concepts modeled and represented by these latent variables.

Most theoretical concepts are conceptually assumed to be quantitative, and measurable on an interval scale, though with an unknown unit of measurement. Therefore, any model with quantitative latent variables is automatically, at least implicitly, a *scaling model*. Here are just a few variants of possible hypothetical scales and measurement units represented by a metric latent variable:

- *Latent z-scores* are used by a model whose latent variables are assumed to take on latent scores on an interval scale measured in hypothetical SDs generated by the model for the given sample (that may possibly be inferred for a population). Of course, then the latent scale for f has a zero mean

and its unit has no real-world dimension. This property is usually built into most cases of LCF models, or it may be prescribed by the researcher. This is convenient especially if the latent variable is intended to model and represent a theoretical concept in a study focused on concept identification.

- *Acceptation of the manifest metric* of one of the indicators is used and recommended by some authors and in some models (e.g. Lisrel, etc.) or in certain circumstances. Then the metric of latent variable *f* is prescribed to be identical with the scale of a certain chosen indicator *x*, which means that *f* is represented in real-world units of measurement (i.e., kg, cm, number of correctly answered items, etc.). It may be motivated by the use of the "best" indicator with its measurement units in practical decisions. Then the measurement scale of this indicator x (an indicator with reasonably high factor validity) may serve as the marker of a given construct (i.e., the latent variable *f*). The same strategy might also be reasonable if a covariance matrix is the object of modeling or, especially, if all manifest variables are measured in the same units (which, however, is a rare case).
- *Logit scale* may be an alternative to *z*-scores, being another version of "dimensionless" units on an interval scale. It comes into consideration in models that are focused on modeling proportions (frequencies), approximating the probabilities $p(x)$, rather than the values of *x*. The name "logit" stands for "*Lo*garithmic Un*it*," and it was incorporated into latent variable modeling by Rasch from its former use in biometrics and quantal experiments. In the Rasch model, which is a special case of the latent ogive model, the logit is the natural logarithm of the ratio of the difficulty of a binary item *x* to its complement, i.e. $\ln(p_x/q_x)$. Though logits are approximately transformable into *z*-scores under certain conditions, it is still very illustrative to understand Rasch's original motivation in building his model.
- *Additional restrictions*, including some that may be *external* to the latent model, can influence the latent scale. For instance, if two latent variables are hypothesized to be in a certain type of nonlinear relationship, such an additional requirement would influence the metric of the latent variables.

As was mentioned in Section §6, it is possible to theorize about concatenation operations on the scale of a latent variable that models a prospective scientific theoretic-conceptual variable such as "ability." It is possible to imagine an ideal experiment, say "concatenation of an increment of ability" due to a learning or training effect. Of course, with the various conceptually different types of theoretical concepts ("constructs"), such a "concatenation", can be difficult to imagine. Nevertheless, for some, it is possible even to imagine an indirect operationalization of the concatenation—as with learning that improves "ability." Hence, it is possible to conceptualize concatenation *not* only as an *empirical* operation, which is the typical case in representational measurement theories. Therefore, I suggest that the term "empirical" be replaced by the more general notion of *"substantive"* concatenation,[6] or *"semantic-content"* concatenation, which includes empirical as well as theoretic-conceptual concatenation. Otherwise, those theoretical concepts, whose improvement is a target of an intentional process, say an educational process, could not be measured and the improvement expressed. It might also seem that concentration using the term "empirical" instead of "substantive" could be an unintended relic of the influence of operationism.

Historically, the terminology has been influenced by specific areas of application of various models although many of them may be shown to be mathematically equivalent. For example, we can

6 Perhaps automatically thinking of concatenation as empirical concatenation rather than the more general term of "substantive concatenation" could be an unintended relic of the era of operationism.

find three different names (and different standard symbols) for the same kind of *metric quantitative latent variable f* whose scale is usually represented in virtual *z*-scores.

5.3.2.3 Latent Factors, Latent Traits, and Latent Abilities

Latent common factor—f, a typical name used for the latent common factor model, the *LCF model*, is applied to the correlation or covariance matrix of interval manifest variables, or more rarely, to a tetrachoric correlation matrix from dichotomous manifest items, or to a polychoric correlation matrix from ordinal polytomous manifest variables. The historical origin of using the term "factor" in this connotation comes from Spearman's and Thurstone's times in the early study of intelligence as a single "General Factor *g*" vs. multiple "Primary Mental Factors." The scaling property of the LCF model is not usually used in most applications, though standard commercial software offers a variety of computational methods to construct latent *factor scores*, that is, the special forms of *constructional definitions* of latent variables. These methods (e.g. the so-called regression method, Bartlett's method, etc.) are acceptable as constructional association functions only in the case when one uses the genuinely falsifiable LCF model, not the principal components/axes computational routines. In addition, a conceptually meaningful application is limited to content homogeneous indicators.

Another subcategory of models with quantitative latent variables belongs to the family mostly known as *Latent Trait models*, which is usually considered to include the subgroup of *IRT models*, the models of IRT, which are generally used to analyze binary or polytomous manifest variables (items). These are appropriate for dichotomous indicators—the traditional (0; 1)—*items*. They also include the Rasch model with a possible logit scale for the latent variable mentioned above. These models are specifically aimed at modeling the proportions $p(x_j)$, not the scores x_j. These are usually more or less explicitly also used for *scaling*, usually with a scale in the form of standardized *z*-scores and/or logits. By tradition, the names such as "trait" or "ability" are being used here for the latent mathematical variable. In the past, say up to the nineteen fifties, this mathematical latent variable used to be symbolized by φ (phi, but do not mistake it for the fourfold phi-coefficient of correlation). More recently, it is typically denoted by θ (theta). Its separated latent values are usually called 'levels of the 'trait' θ' or 'levels of 'ability' θ.' Of course, 'ability' and 'trait' are not used in the psychological sense, but as purely mathematical technical terms—commonly used labels for a statistical latent variable of a special kind, namely the metric latent variable used in IRT models for dichotomous items x.

The *"latent trait" (θ)* is a typical term used in modern *IRT modeling*. The name is in quite general use in that group of models and applications. Sometimes there is an interesting mismatch between the mathematical variable called "trait" and the corresponding psychological "trait." For instance, in personality theories, we find "traits of personality." However, one of the leading theories is Cattell's 16-PF—Personal *Factors*, not "Traits." In addition, in those social science areas that work with such theoretical concepts as psychological/behavioral "traits," the indicators need not be just dichotomous, which is the specialty of the so-called "latent trait" models.

"Latent ability" is a name that has been mostly used in works dealing with older attempts to elaborate *latent ogive models*, which is more recently included as part of the family of IRT models. They were substituted in place of the crude methods of the so-called Item Analysis. In those older writings, instead of *f* or θ, the authors used the symbol φ (phi) for the "latent ability" variable (again, this φ has nothing to do with the phi-coefficient of fourfold correlation). It is a matter of habit or taste which symbols are preferred in the literature. Usually, *f* is being used in LCF models for a latent

common factor, while θ is typical in IRT models (the symbol φ is not frequently used in recent literature). Recent mathematical derivations based on the mutual relationship $p(x|f) = E(x|f)$ show that there is a simple straightforward mutual relationship between f and θ so that, in general, f might be used in a broader sense for any type of latent variable.

Again, we can see some contradictions between the mathematical variable "ability" and the behavioral concept of "ability." The older applications of the latent ogive and Rasch models were mostly applied to the measurement of a student's "academic ability" by a battery of binary items; therefore, the name "ability" was used also for the mathematical variable. At present, however, the term "latent trait analysis" is generally preferred; therefore, we could read a combination like the "level of children's reading ability was modeled by a latent trait model" and so on.

Ordinal or so-called *nonparametric* models. There could be a case when manifest variables are dichotomous items and the latent variable is assumed to have ordered polytomous categories. The applications of this model are not too frequent these days. In addition, there are some models with continuous latent variables but with constructional association functions that are very weakly specified, mostly required only to be monotone, i.e. preserving rank order. The so-called Mokken's scale model is of that type. Again, applications of these models are not too frequent nowadays.

5.3.2.4 Latent Class Models

The *Latent Class* model for dichotomous items is based on the idea that the theoretical concept(s) may be purely qualitative, i.e. of the nominal type. Hence, they are not orderable as in the models above. It can be used preferably for testing or developing a qualitative typology, say sorting people into *latent typological classes*. Its special case is the *Latent Dichotomy* model with only two latent classes, which can be formally coded

$$f = \left\langle {1 \atop 0} \right.$$

These models were developed by Lazarsfeld after the Second World War with applications to typologies of people in sociological research.

For some time, Lazarsfeld's original development had been running historically in parallel with factor analysis, which was already in a more elaborate form back then. His conception covered a collection of certain latent models under the common name of *Latent Structure Analysis* (*LSA*). It included primarily the Latent Class and Latent Trait models including models similar to Latent Ogive and some other IRT models (to use the up-to-date terms). In addition, a special terminology was in use, which may still be helpful for understanding the bi-directional character of association functions in today's world:

- "*Accounting equations*" correspond to the mathematical *modeling derivations* or to the latent *measurement model* equation (if the latent variable were quantitative). In LSA, these equations were "accounting for" the observed proportions/probabilities, including the joint proportions/probabilities, by their conditional probabilistic dependence on the latent classes, $p(x|f)$ or $p(x_j, x_k|f)$, ... In the quantitative case, they would correspond to the *measurement version of the association function*. In the LCF model, its counterpart is the factor model, i.e. a regression equation with the LC factors as independent, and the manifest variables as dependent variables.
- "*Recruiting probabilities*" correspond to the *constructional definitions*. They assign a subject to an appropriate typological latent class according to her/his conditional probabilistic dependence on successfully passing a certain subset of the battery of dichotomous items. For example, we may

imagine the probability that an individual is "recruited" from a certain latent class if we know his/her given profile or pattern of responses on n items x_j, schematically $p(f|x_1, ..., x_j, ..., x_n)$.

In the quantitative case, they would be related to the *construction version of the association function*, in the LCF model, their counterparts are formulas for estimation of latent factor scores.

5.3.2.5 General Latent Variable Modeling

After the formal mathematical relationships among these and other models had been discovered (McDonald, 1962), it was necessary to distinguish among the various types of these models that could be mathematically reformulated into a general system of regression equations. This enabled a unified mathematical formulation of a rather wide class of such models, making separate terminologies unnecessary. However, due to tradition and some narrower conceptions of certain types of practical applications, different names for mathematically identical things are still being used.

There is an open question with respect to how well a certain type of probability distribution function may be based on conceptual theory in an area of research where latent variable models are used. In most cases, the answer rests simply on "what is at our disposal in the software." Typically, it is the case that the *normal distribution* is rather automatically used in latent structure models, including IRT, as well as in LCF models. Nevertheless, currently, one can use *distribution-free* models, but those typically need huge sample sizes, say N to be at least several hundreds or even several thousands, if the number of manifest variables is large. Therefore, very often not only a *normally distributed marginal* distribution but also a *multivariate normal* joint distribution is being assumed, and somehow attempted to be approximated. In practice, a typical approach in LCF models consists of computing and using tetrachoric correlations for dichotomous items, and *polychoric* correlations for ordinal categorical items as the basis for model fitting. This means that the measuring scales of pairs of manifest variables are forced to approximate a two-dimensional normal distribution, but such scales are neither likely to stay the same for all combinations of one variable with the n-1 others nor do they guarantee that the pairwise normality would produce overall joint multinormality. Most models are fitted on such a basis, but their acceptability should be judged rather as a general approximation than a genuine statistical test.

In addition, in some IRT latent trait models for dichotomous indicators (items) with the latent variable usually technically called the "latent trait," it is assumed that the observed proportions $p(x_j)$ of n items are manifestations of the latent variable, the "trait," *with a normal probability distribution*—as a theoretically based conceptual presumption about a human attribute. However, recently, some IRT models do not require such an assumption. Therefore, it is not clear whether the assumption should still be used as a substantive-conceptual necessity.

Especially, in linear LCF models, if multivariate normality of the manifest as well as the latent variables is assumed, the description of the probability distribution is completely defined through the *first-* and *second-order moments*, including mixed moments. It means that the variances and covariances (correlations) are sufficient to generate the probability function. Thus, higher moments, such as the marginal third order for asymmetry, and/or higher mixed moments for nonlinearity, etc., are not needed. But, on the other hand, if there were a conceptual substantive reason, the researcher might insist that he/she is going to model the properties of the data just up to the information[7] described by second moments, leaving

7 The above-mentioned term "information" has been used here in a generic sense. It has nothing to do with Fisher's mathematical *information* model related to error variance of estimation or with Shannon's *information and entropy* both described in Section §46.

the rest as conceptually unimportant or uninteresting. Therefore, the numerical analysis is usually applied to modeling the values, observed scores, and their covariance/correlation matrices.

In IRT modeling of item proportions, the complexity of the modeled objects usually does not exceed the marginal $p(x_j)$ plus two-dimensional joint proportions $p(x_j, x_k)$. The latter may be organized in a kind of two-dimensional contingency table, an analog of a correlation matrix. By analogy of modeling third- and higher-order moments of x to describe asymmetry, nonlinearity, etc., it would be possible to model three-dimensional joint proportions for all possible triples of items organized in a cube, or possibly four-dimensional for quadruples, etc., organized in a four-dimensional contingency table. Only theoretically, if all possible combinations of all n items organized in an n-dimensional cube were used as an object of modeling and as the source for computing the parameters of the latent model, this would be called a *full-information* method of estimation of the model parameters, while a *partial-information* method would involve a combination of less than n items. But applications to more than two-dimensional joint proportions (approximating probabilities) are very rare. Most of us just remember one or two applications to three-dimensional joint relative frequencies.

5.3.2.6 Uncorrelatedness and Probabilistic Local Independence

The type of *local probability independence* used is also one of the important conditions that determine the character of the latent variable model. In Section §35, §36, and §37, we already explained how local correlational independence works as a metaphor for an explanatory principle. There are two ways to look at the variety of types of specification of it.

1) From one point of view, we may distinguish between:
   a) The *stronger* version—local correlational independence, meaning partial uncorrelatedness, i.e. linear regressive independence, which asks that the manifest *correlations* be "explained" by the latent variables. This is typically applied in LCF models and linear structural models.
   b) The *weaker* version—local probabilistic independence, i.e. conditional probability independence, when complete probabilistic independence including nonlinear and more general kinds of relationships can be "explained." However, this version is weaker mathematically, as it does not impose a restriction of linearity. It is more rigorous from the point of view of application and interpretation since it asks for a full explanation of even complicated stochastic relationships. It is typically used in IRT and other LSA models that deal with joint probabilities $p(x_j, x_k)$ rather than with correlations.
2) The possibility of excessive liberalization of the principle leads to another classification:
   a) *Complete* (*full, absolute*) local independence allows no kind of liberalization, it means that the manifest correlations (in 1a) and or probabilities (in 1b) have to be "explained" fully, up to zero, as far as possible.
   b) *Structured and patterned* local independence represents liberalization of (i) in certain special cases such as longitudinal LCF, inter-battery LCF, and so on. In these special versions of latent variable models, some manifest correlations or proportions are intentionally prescribed in advance not to be fully "explained."

5.4 Summary

While Content Validity is first in Chapter 5's title, it is best to summarize it following a review of key points related to this chapter's treatment of IRT. Petr Blahuš spoke of IRT frequently but did not anticipate how strongly IRT was going to move into both scientific and practical work in psychometrics. It was unthinkable to place Test Theory in the title of this book without giving the reader the ability to perform IRT analysis, CTT, and latent common factor analyses. The most important element in this book to help the reader become proficient with IRT concepts and methods is Computational Guide #2. It will prove to be more than worth your time. The same is true of Computational Guide #3 on factor analysis.

One weakness of seeking latent factors through factor analytic methods is that the interpretation of the factors is primarily through an understanding of what the tasks demand. This is natural because the number of tasks or items and their groupings is much smaller than the number of people and the way people are ordered in proficiency from lowest to highest on a single latent factor dimension. This feature of factor analysis and related methods may have been one reason Sam Messick, in describing domain theory, emphasized "construct-relevant sources of task difficulty." But the model that provides explanation of what the test-takers are doing is what leads to the development of construct-relevant tasks.

This is called the person or student model in ECD. It leads directly to the task models. In other words, an understanding of what persons must do cognitively and affectively should lead to the development of tasks. It is unwise to assume that items in existing item banks can be found that include the kind of mental processing required for criterion attainment in the selected domain. In this chapter, we have tried to show that it is the domain-specific model of progressive attainments of people ranging from low to high proficiency that provides important interpretive information. Ostensive demonstration with ratings of the constructs may be superior to tests constructed from pre-existing pools of conventional items in providing meaning and explanation about levels of progressive attainment. This idea of the ordering of differences in levels of strength in using a trait, or a practiced habit, applies across psychological and medical measures, wherever humans can manifest practiced attainments. Messick used the term "domain theory" with his task emphasis, yet the "person model" is the essential element in guiding the construction of tasks. ECD and domain-specific model-based reasoning properly list the person model before the task models.

The Wright map discussed in §52b puts the person distribution and the item difficulty levels on the same scale. This is what IRT does with greater accuracy of placement than does CTT. Latent Factor extraction does not have a convenient way to calculate a value for each person on the factor. Factor scores are feasible, but because of the variable of how the factors are rotated for a final solution, comparability from one session to another is hampered.

In the McBride study of FORE, the hypothesized latent factors of Word Knowledge and FORE were determined first, and then IRT was used to order the constructs of phrase boundaries, smoothness, pace, confidence, and expression. Note that this sequence of progressively difficult attainments shifts from the cognitive domain for the first three, then conative and affective for the last two. Certainly, these are not homogeneous except the ratings were all from a standardized ostensive performance, corrected for the nonfocal but correlated word knowledge dimension. Arriving at the right set of rating scales in this model of progressive attainment was benefitted from a unique family focus on elementary reading, augmented by literature reviews so that a sensible series of rating scales could eventually perform as a unified FORE factor. A teacher story was developed after the study to enable reading teachers to provide feedback after each student read out loud. The study did not have funding to continue into a study to evaluate the effectiveness of repeated

reading sessions over the four years of elementary school, but the interpretive information and rating scales are there for such follow-up.

Here are some points we leave with you in summary:

1) Latent Factor studies are needed initially to establish the dimensions in a domain of "content" for a particular population of people. Single dimensions are needed for the local independence required for quantitative scales at the interval level. Theory must guide the development of instruments, utilizing tools such as rating scales and ostensive demonstrations, which are often underused. These tools help uncover underlying person models and develop tasks that genuinely align with the measurand, particularly in learning situations aimed at fostering progressive attainment. This principle applies in health progression as well, into and out of disease states with one kind of therapy (and medical testing) initially as the disease is discovered and brought under control and other therapies for healing and rehabilitation are undertaken.
2) IRT studies follow up, with all latent factors being used in study designs for control of extraneous dimensions, but with a focus on the latent variable(s) of greatest interest.
3) The explanation of the levels of progression for various domains, such as learning, therapy, sports-performing arts, or many other areas within the human sciences, may not have "homogeneous content." Instead, quite different cognitive, conative, or physical processes may occur at each level. This underlying reality has a tendency to produce "difficulty factors" where you have reason to believe the construct is a complex but unified performance. Fixing this problem may require changes in the rating scales or other tasks, or in the instruction at the time the assessment is performed. IRT, once you know what characterizes a single dimension, is very effective and powerful and provides explanatory power when the instrument has been developed using good models—including especially a good model of person progression.

5.4.1 Pitfalls to Avoid in So-called *Content Validation*

Expert judgment is not the only "common law practice" that should be undertaken with caution. Another is the test blueprint. Recognize that it is the polar opposite of the vague but sensible notion of content homogeneity. It mixes diverse dimensions up in the service of state or professional mandates. Such tests or evaluations meet these mandates as required, and wise educators use them and the time they take as best they can to help learners. Omelets of scrambled content serve the mandates, but they are *NOT* learner-centered *NOR* a part of metrology *NOR* a solution to the content aspect of validity for tests designed scientifically to uncover and measure essential dimensions of growth. Rather they try to balance and manage the content check-off process with the science of the evaluation of competencies if well done.

5.4.2 Latent Variables and Factors

The last section of the chapter was like a guided tour, a somewhat historical guided tour, through the wide variety of latent variable models ranging from the latent common factor to latent traits, to latent class models, to latent abilities, and even a brief discussion of McDonald's derivation of a mathematically generalized latent class model. In summary, Petr has taken a few more steps into the latent variables and latent factors story. We will move into it more deeply in the next two chapters.

5.5 Study Questions

5.1 Design and Initial Validation of Latent Factors and Domain Models

Q1 Summarize the steps in Reo McBride's dissertation study of fluent oral reading. (Section §49)

Q2 What is domain theory? What is evidence-centered design? What is a domain theory of progressive attainment? *(Subsection 5.1.1.3 and Subsection 5.1.1.4).*

5.2 Content Validity and Item Response Theory

Q3 Explain the difference between "declarative knowledge" and "the substantive cognitive processes crucial to criterion performance." *(Subsection 5.2.1.1)*

Q4 Which is more difficult to acquire, "declarative knowledge," or "the substantive cognitive processes crucial to criterion performance? *(Subsection 5.2.1.1)*".

Q5 Why are questions not the best way of measuring the skills of performance, that is, the procedural knowledge of skills? *(Subsection 5.2.1.1)*

Q6 Compare and contrast "declarative linguistic knowledge" (know-what) with a skilled performance (know-how). What is the best way to assess each of them? *(Subsection 5.2.1.1)*

Q7 Comment on the statement in the first paragraph under "causal explanations," that Borsboom et al. (2004) "... regard causation as the simple minimal core of validity." In what ways is it the minimal core concept of validity? *(Subsection 5.2.1.2)*

Q8 How is that "minimal core" connected to the logic of establishing a latent variable using the latent common factor model? *(Subsection 5.2.1.2)*

5.2.4 Content Validity and Content Homogeneity of Indicators/Subtests

Q9 What are some of the indices for measuring "expert agreement"? What are the strengths and weaknesses of "expert agreement" as evidence for content validity? *(Subsection 5.2.3.2)*

Q10 What is content homogeneity? What are some of the other types of homogeneity with which it may be confused? *(Subsection 5.2.3.3)*

Q11 Compare and contrast content homogeneity with unidimensionality. *(Subsection 5.2.3.4)*

Q12 Comment on the sentence that "from one point of view, a cluster of indicators found to be content homogeneous need not be unidimensional in terms of its mathematical model." How could this be explained by "difficulty factors"? *(Subsection 5.2.3.4 and Subsection 5.2.3.5)*

5.2.5 Some Aspects of Item Response Theory and Computer-Adaptive Testing

Q13 What is *Item Response Theory* (IRT) and how it is related to computer-adaptive testing? *(Section §52 and Subsection 5.2.3.4)*

Q14 What are Item Characteristic Curves (ICCs) and how are they used in IRT? How are they used in Wright Maps? *(Subsection 5.2.3.1 and Subsection 5.2.3.2)*

Q15 Explain the connection between IRT and CAT as exemplified in their application to the analysis of the skill of high jumping. *(Subsection 5.2.3.3 and Subsection 5.2.3.4)*

Q16 Ho does the high jumping application domain emphasize some of the important differences between "classical tests" and "computer adaptive tests." How do these relate to the differences between the Latent Factor Model and *Theory? (Section 5.2.3.3)*

5.3 Latent Variable Modeling

§53 Generic Score of Homogeneous Indicators: Rough Approximations Without Latent Variables

Q17 Summarize the historical origin of the content homogeneity concept with Jane Loevinger in the 1940s and the value of that concept in contemporary psychometrics. *(Section §53)*

Q18 What are some of the weaknesses and strengths of various types of "formula scores" employed to ostensibly create reasonable estimates of scores for a latent Chapter division **5.3**, first two paragraphs. *(Subsection 5.2.1.1)*

Q19 Summarize some of the problems and shortcomings of so-called "factor score" methods, even some that are institutionalized in commercial statistical software. *(Subsection 5.2.1.1 and Subsection 5.3.1.3)*".

§54 The Types of Manifest and Latent Variables in Latent Variable Modeling: Values, Scales, and Probability Distributions

Q20 How is the latent variable modeling process affected by the types of latent and manifest variables being modeled? *(Subsection 5.3.2.1)*".

Q21 How is latent variable modeling affected by the level of quantification chosen for latent and manifest variables? *(Subsection §54).*

Q22 What are latent class models? *(Subsection 5.3.2.4).*

Q23 Briefly describe the historical development of Lazarfeld's Latent Structure Analysis during the 1940s and its relation to contemporary latent class models. *(Subsection 5.3.2.4).*

Q24 Give a brief account of McDonald's mathematical synthesis of the various approaches to latent variable modeling toward a unified formulation. *(Subsection 5.2.3.3, Subsection 5.3.1.2, footnote 5, Subsection 5.3.2.5).*

References

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications*. Information Age Publishing Inc.

Bunderson, C. V. (2011). Developing a domain theory defining and exemplifying a learning theory of progressive attainments. *Journal of Applied Measurement*, *12*, 25–48.

Bunderson, C. V., Wiley, D. A., & McBride, R. (2009). Domain theory for instruction: mapping attainments to enable learner-centered education. *Instructional-design Theories and Models: Building a Common Knowledge Base*, *3*, 327–347.

Brennan, P. F., & Hays, B. J. (1992). The kappa statistic for establishing interrater reliability in the secondary analysis of qualitative clinical data. *Research in Nursing and Health*, *15*, 153–158.

Cicchetti, D. V. (1984). On a model for assessing the security of infantile attachment: issues of observer reliability and validity. *Behavioral and Brain Sciences*, *7*, 149–150.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Coderre, S., Woloschuk, W., & McLaughlin, K. (2009). Twelve tips for blueprinting. *Medical Teacher*, *31* (4), 322–322.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement, 2nd edition*. Washington (DC): American Council on Education, pp. 443–507.

Downing S. M., & Haladyna T. M. (Eds.). (2006). *Handbook of test development*. Lawrence Erlbaum Associates.

Downing, S. M., & Haladyna, T. M. (Eds.). (2011). Content-related validity evidence in test development. Handbook of Test Development. https://doi.org/10.4324/9780203874776-13

Embretson, S. E. (2007). Construct validity: a universal validity system or just another test evaluation procedure? *Educational Researcher 36*, 449–455.

Embretson S. E. (2010). Measuring psychological constructs with model-based approaches: an introduction. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches*. Washington, DC: American Psychological Association, pp. 1–7.

Garner, M. L., Engelhard, G., Fisher, W. P., & Wilson, M. (2011). Advances in Rasch measurement: Volume 1. JAM Press.

Kane, M. (2006). Content-related validity evidence in test development. In S.M. Downing, & T.M. Haladyna (Eds.), *Handbook of test development*. Lawrence Erlbaum Associates, pp. 131–153.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437–448.

Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions and applications*. Information Age Publishing.

Lord, F. M. (1950). Properties of test scores expressed as functions of the item parameters. *ETS Research Bulletin Series*, *1950*(2), i–8.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, *61*(4) (Whole No. 285).

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, *45*, 507–529.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*, 382–385.

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences*. Springer Series in Measurement Science and Technology. Springer.

McDonald, R. P. (1962). A note on the derivation of the general latent class model. *Psychometrika*, *27*, 203–206.

McDonald R. P. (1999). *Test theory a unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. CRESST Report 752. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mulaik, S. (2005). Looking back on the indeterminacy controversies in factor analysis. In A. Maydeu-Olivares, & J. J. McArdle (Eds.), *Contemporary psychometrics, a festschrift for Roderick P. McDonald*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers, pp. 173–206.

Mulaik, S., & McDonald, R. P. (1978). The effect of additional variables on factor indeterminacy in models with a single common factor. *Psychometrika*, *43*, 177–192.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. V. (2010). The Rasch model and additive conjoint measurement. In Garner, M. L., Engelhard, G., Fisher, W. P., and Wilson, M.*Advances in Rasch measurement, Volume 1*. JAM Press.

Raymond, M. R., & Grande, J. P. (2019). A practical guide to test blueprinting. *Medical Teacher*, *41*(8), 854–861.

Risconscente, M. M., Mislevy, R. J., & Corrigan, S. (2006). Evidence-centered design. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development*. Lawrence Erlbaum Associates, pp. 40–63.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., Shwartz, Y., Hug, B., Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, *46*, 632–654.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences, 2nd edition*. New York: McGraw-Hill.

Standards for Educational and Psychological Testing. (1974). *American Psychological Association (APA), American Educational Research Association (AERA), National Council on Measurement in Education (NCME)*. Washington, DC: APA.

Standards for Educational and Psychological Testing. (1999). *American Psychological Association (APA), American Educational Research Association (AERA), National Council on Measurement in Education (NCME)*. Washington, DC: APA.

Vernon, P. E. (1971). *The structure of human abilities*. London: Methuen.

Wright, B. D., & Master, G. N. (1981). *Rating scale analysis*. Chicago, IL: MESA Press.

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research*, *25*, 508–518.

Zalles, D., Haertel, G., & Mislevy, R. J. (2010). *Using evidence-centered design to support assessment, design, and validation of learning progressions* (No. 10). PADI Technical Report.

Handbook of test development. https://fatihegitim.files.wordpress.com/2014/03/hndb-t-devt.pdf

6

The Latent Common Factor Model

CHAPTER MENU

6.1 Latent Variables and the Common Factor Model, 226
- 6.1.1 (§55) Representation Conditions of Measurement and Correspondence Rules for Latent Variable Modeling, 226
  - 6.1.1.1 Semantic Correspondence Rules, 226
  - 6.1.1.2 Syntactic Correspondence Rules, 227
- 6.1.2 (§56) The Latent Common Factor Prototype of Latent Variable Modeling, 230
- 6.1.3 (§57) How to Calculate Loadings and Residual Correlations in the Factoring Process, 232
  - 6.1.3.1 Applying the Latent Common Factor Model, 232
  - 6.1.3.2 Centroid Method of Factoring, 234
  - 6.1.3.3 Comparison of Three Models: The "Merely Historical Centroid Model," the PC Model, and the ML Model, 235
  - 6.1.3.4 Fundamental Factor Theorem, 238
- 6.1.4 (§58) Maximize the Fit of the Model and Minimize the Discrepancy Index to Achieve Local Independence, 239
  - 6.1.4.1 Least Squares Fitting and Model Fit Assessment, 239
  - 6.1.4.2 Relative Model Fit, 240
  - 6.1.4.3 Unidimensionality and Content Homogeneity, 241
  - 6.1.4.4 Weighted Least Squares (WLS) Fitting, 241
  - 6.1.4.5 Maximum Likelihood (ML) Fitting, 242

6.2 Factor Scores, 243
- 6.2.1 (§59) Estimating Each Subject's Latent Score *f* to Measure a Trait or Characteristic, 243
  - 6.2.1.1 Estimating Factor Scores, 244
  - 6.2.1.2 Factor Scores and Latent Values, 246
  - 6.2.1.3 Generic True Scores and the Latent Common Factor Model, 247
- 6.2.2 (§60) Communality, the Variance Shared Among Indicators, 247
  - 6.2.2.1 Communality, Specificity, and Unreliability, 248
  - 6.2.2.2 Communality and the Squared Multiple Correlation (SMC) with the Other Indicators, 249
- 6.2.3 (§61) An Example Regression Estimate of *f* and the Decomposition of Indicator Variance, 250
  - 6.2.3.1 The Tapping Test, 251
  - 6.2.3.2 Test–Retest Analysis, 251
- 6.2.4 (§62) Generic Reliability and the "Model–Remodel" Assessment, 253

6.3 Summary, 256
6.4 Study Questions, 257
References, 274

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

6.1 Latent Variables and the Common Factor Model

The first part of this chapter presents the logic, mathematical models, and quantitative basis for several approaches to the calculation of *factor loadings* from matrices of correlation coefficients or covariances among indicator variables. These methods are used to show how indicators—tasks or ratings—enable the hypothesized factor to be confirmed, along with other factors if there is more than one. The second part of the chapter deals with completing the process by creating *factor scores* for persons on the obtained latent variable(s).

6.1.1 (§55) Representation Conditions of Measurement and Correspondence Rules for Latent Variable Modeling

Representation conditions generally refer to nominal, ordinal, interval, or ratio scales. The selection of a model that delivers a scale at one of these levels means that the latent variable produced by the model will deliver scale values corresponding to that level of representation. Mathematical latent variables (e.g. f, θ) work as model counterparts of theoretical concepts. The nature of the theoretical concepts, after they have been clearly formulated as conceptual scientific variables, also has to be made clear with respect to their quantitative or qualitative character. Nominal and ordinal scales are qualitative, but ratio scales are rare in the human sciences, leaving interval scales as usually the highest level available. Therefore, an appropriate latent model has to be chosen corresponding to the level of the measurement scale of its latent variables, and according to the theoretical background of the theoretical-conceptual variable. In this sense, the type of latent variable of a certain mathematical model determines the representation conditions for the "level of latent measurement." In the indicated way, the theoretical-numerical representation produces the numeric values of the mathematical latent variable as a stochastic-random latent variable through the process of modeling. In that not-directly observable layer, it works as an analog to the representation conditions for the empirical numerical representation of the observed empirical research variable, which then takes on the form of a stochastic-random manifest variable through the operation of a measurement instrument in obtaining raw scores for indicators. Fulfilling the empirical representation conditions determines the corresponding level of direct measurement of the manifest variable. These two analogs are presented in Figure 6.1.

6.1.1.1 Semantic Correspondence Rules

An additional aspect deals with the mutual relationships between the empirical and theoretical levels. The empirical attributes and research variables work as observable indicators of the unobservable theoretical concepts ("constructs"). Their mutual relations are known as correspondence rules. In the not-yet-formalized, nonmathematical stage of progress in science, these are more or less expressed by their substantive-semantic content, usually formulated in a verbal form. The rules of correspondence are bi-directional consisting of definitions and interpretations.

A theoretical concept is formulated by its semantic definitions, also referred to as "epistemic definitions" (Torgerson, 1958; Lord and Novick, 1968). The theoretical concept is defined in terms of its empirical indicators, however only partially since the list of all possible indicators hardly could be complete. The semantic definitions of the substantive content of theoretical concepts are also called "constitutive" or "construction rules" (Maraun, 1996; Guttman, 1955). For instance, the semantic content of "sensorimotor coordination ability" is constituted in terms of a cluster of "content-homogeneous" coordination tests that are practically executable under standardized conditions yielding registerable test scores.

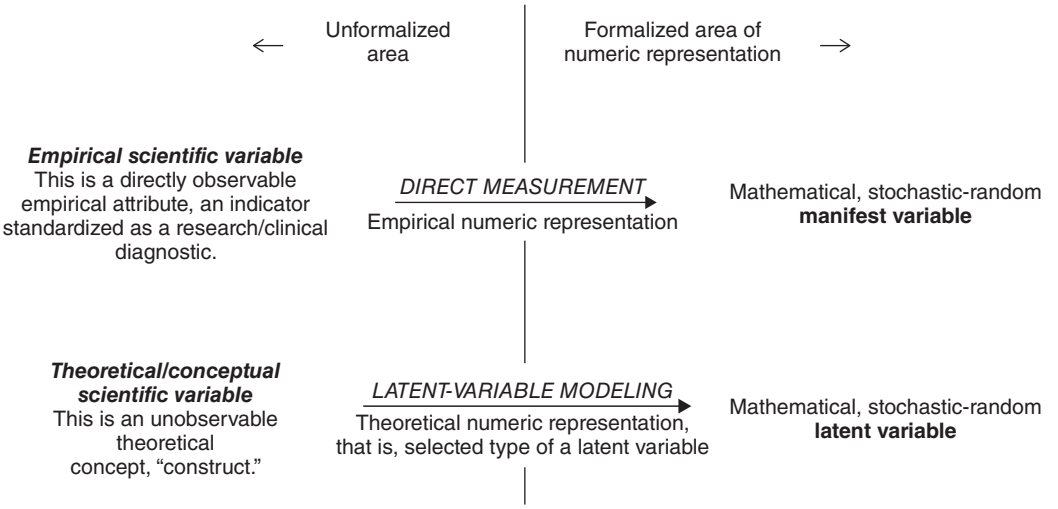


Figure 6.1 A representational measurement theory point of view applied to manifest and latent variables.

In the opposite direction, a theoretical concept gets the empirical semantic interpretation of its theoretical content through the indicators. The number of indicators used in a practical situation is limited. In addition, an empirical manifestation of the theoretical concept is always joined with certain specific features of a given indicator such that an interpretation is necessarily only partial. In Hempel's (1966) terminology, these partial interpretations are called interpretative sentences. For example, a practical projection of "physical coordination ability" is through observable performance in a cluster of coordination tests. However, each test contributes to the interpretation of "what coordination ability is like" from the position of the concrete formulation of its test tasks, and the corresponding specific content of the movement actions for its performance.

6.1.1.2 Syntactic Correspondence Rules

The above-mentioned semantic correspondence rules have their *syntactic counterparts*, i.e. the former become *formalized*, quite often in terms of a *mathematical function*, typically in the form of a *regression function*,[1] as one of many types of mathematical models of latent variables.

A semantic definition is now modeled by a formal *syntactic definition* that gets a form of mathematical construction of estimates of the latent variable with its latent values in subjects as shown in Figure 6.2. In its mathematical form, the definition may be also called a *construction formula* (term by Guttman, 1955). For instance, in the LCF model, the method of construction is expressed by formulas for computing the z-scores of f, the latent factor scores of subjects (such as the well-known Bartlett's or Thomson's formulas). In IRT models, the latent z-score values of θ are usually called levels of an "ability" or of a "trait."

1 Some authors do not distinguish between semantic and syntactic aspects, automatically considering the indicators as well as theoretical concepts directly as quantitative variables and their mutual relationships straightforwardly just as mathematical functions. Moreover, some of the authors, in their formalistic interpretation, even reduce the general notion of *operational/epistemic definition* of a theoretical concept only to linear regression functions. For example, Costner (1969) called the linear regression coefficients between empirical indicators and theoretical concepts "*epistemic coefficients*." On the other hand, his "*abstract coefficients*" correspond to regressions between mutually inter-relating theoretical concepts (e.g. regressions among latent factors). Such a direct reduction might be too simplistic.

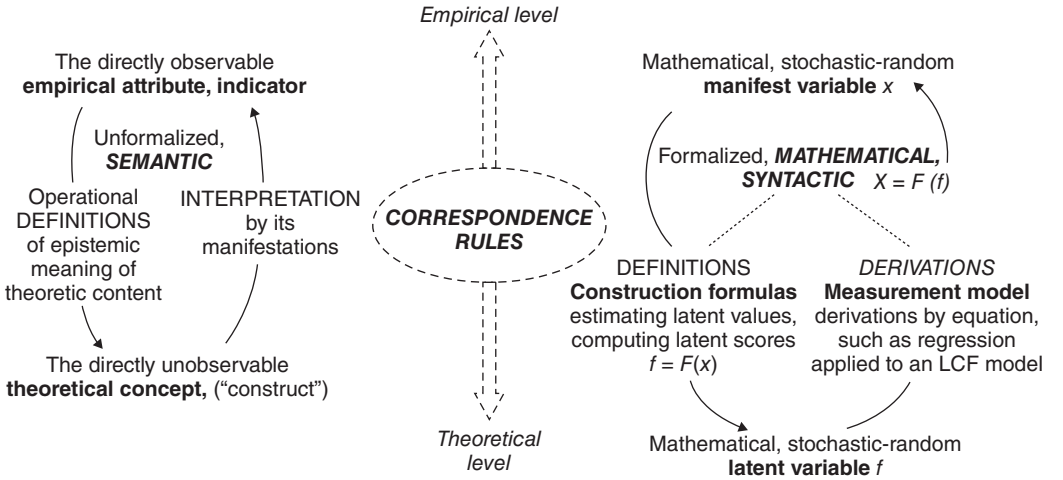


Figure 6.2 Correspondence rules between the empirical and theoretical levels. The unformalized semantic-substantive empirical structure is on the left. Its formalized syntactic reflection is on the right, often as a mathematical LCF model.

Hence, in latent variable modeling, the semantic definition gets the form of a constructive *syntactic definition*, the *construction* version of the *association function*, such as $\dot{f} = F(x)$. The choice of a syntactic definition has to be based on its conceptual appropriateness to express the substantive-semantic content in terms of a mathematical model. In the construction version of the association function created in the Manifest → Latent direction, the observable x indicators work as independent variables with the latent $\dot{f}$ as the dependent variable. If we consider the Gauss measurement-errors model as an oversimplified illustrative parallel example, then the indicated construction takes on the form of the arithmetic mean of replications.

The syntactic counterpart to the semantic empirical interpretation of a theoretical concept takes on the form of a hypothetically prescribed syntactic derivation of the observed and formally "coded data" from the mathematical latent variables. This is carried out in terms of a latent-variable model in the Latent → Manifest direction, that is, by the measurement version of an association function, $x = F(f) + d$, specified in a basic equation as a "measurement model." This model is typically a system *regression equations* $\hat{x} = F(f)$, with f (the latent variables) used as the independent variables, while $\hat{x}$ (the regression estimates of the manifest variables) are the dependent variables. In the illustrative example with the Gauss model as a counterpart, this model would take on its very special and simplified form, $x_j = \tau + \varepsilon_j$.

When latent variable modeling is used, the two features, namely the *representation measurement conditions* together with *correspondence rules* are jointly present and working together as a *methodological framework for measuring theoretical concepts*, that is, measurement in a more general sense. Its scheme, joining the content of Figures 6.1 and 6.2, is presented in an abbreviated form in Figure 6.3.

Professor Blahuš's figure, now designated Figure 6.3, is very important in his logical and conceptual model-based reasoning. It is a part of his quest for a *measurement approach* to latent variable modeling. He thought "Weak Associative Measurement" was the solution, and in this figure, it was the sole entry on a diagonal line going from the upper left to the mathematical latent variable in lower right. However, given recent developments in the science of International Metrology, we have shrunk the title "Weak Associative Measurement" to a smaller font and now indicate it has been fully replaced by three instrument processes. Remember our dialectic with the senior

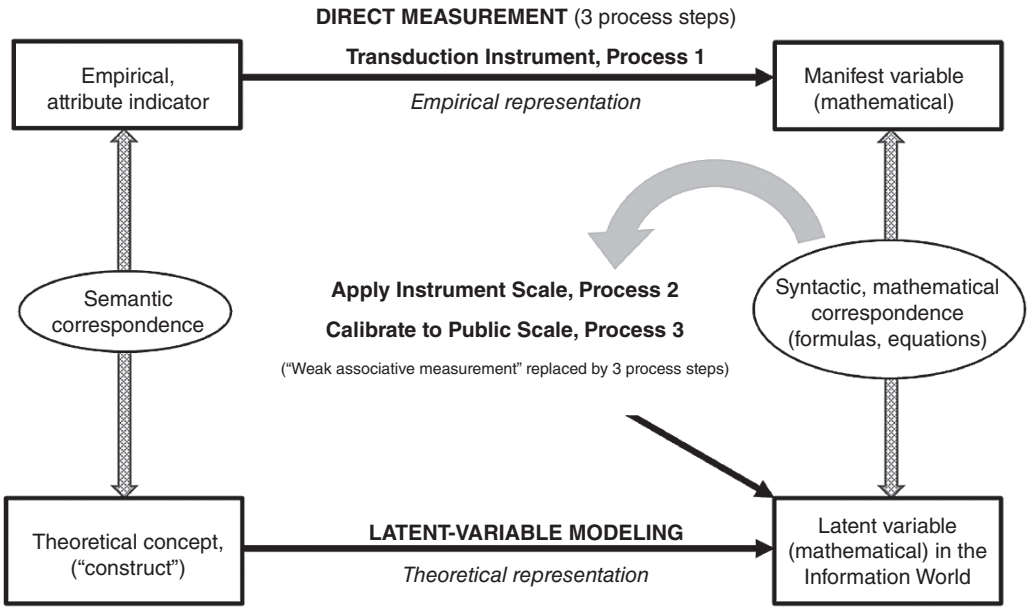


Figure 6.3 Latent variable modeling as a methodological tool combining two interconnections: (i) The attribute–variable correspondence is in the horizontal direction (→) and (ii) the empirical–theoretical correspondence is in the vertical direction (↓↑). Both lead to the Information World.

author mentioned in the preface of this book, about Associative Measurement? We have rejected Associative Measurement and removed discussion of it almost everywhere. But this figure is the ideal place to show how it has been replaced in his well-reasoned and complex conceptual scheme without damaging the scientific logic he was driving toward. This figure is interesting because its top line acts the same as the transduction function, as far as he goes with it, and he derived a calibration function in an earlier section. He does not get credit for two out of three because he did not clarify how transduction worked in all cases, did not include the second process, and did not put together and justify the entire new definition of Direct Measurement we are using. He assumed Weak Associative Measurement was a form of indirect measurement.

But despite this, his logic leading up to this figure about the scientific and mathematical foundations of latent variables is instantiated with the new addition of three Instrument Processes, and the term Information World as the place where latent variables are received, interpreted, and used.

The three designed processes that take place in the instrument and make up Direct Measurement are numbered 1, 2, and 3 in Figure 6.3. Petr designated the first process, the top line going from left to right, as Direct Measurement. He described it as a collection of the raw scores that then need equations. But Direct Measurement is not just restricted to getting the raw scores from the administration of the instrument as indicated on his top line. The full metrological definition of Direct Measurement in Mari, Wilson, and Maul (2021) includes three sequential processes. The first, Transduction, takes the raw scores in and begins to convert them to a scale value using formulas and equations.

We have fully replaced "Weak Associative Measurement" with this transduction process plus two additional intentionally designed processes, The second one, to apply the instrument scale, takes the raw scores part-way through and just into the Information World, as the instrument scale value is assigned. If we take the example of the Fluent Oral Reading (FORE) latent variable, Process 1, transduction takes the raw scores of a person on the FORE rating scales and recalls the previous

estimated parameters for these tasks from the instrument's memory. It calculates the person's IRT score on the logit scale based on these prior parameters. This creates the instrument scale, and that is the end of it if there has been no public scale agreed upon for FORE. If there were, the value on the instrument scale would locate the calibration values in a linear equation to put the value on the public scale before reporting it out to the information world. Note that the curved arrow pointing to instrument processes 2 and 3 comes from the elliptical figure at the center right where the calculations are performed. The manifest variables, that are the output from the Transduction process (Process 1), now undergo mathematical processing to Apply Instrument Scale (Process 2) and to Calibrate to Public Scale (Process 3) if such has been included.

The example of FORE was used because it is complex and awkward to compute a factor score for an individual, and there are factor invariance issues. It is common these days and simpler to use IRT for the latent variable, after understanding the factor structure of the domain variables. If a variable were being treated as a latent factor, another calculation, such as the Bartlett (1938) method presented in Section §55, would have to be completed in advance, and values placed in the instrument, to calculate the persons' factor scores on the instrument scale. In this case the result is still the *instrument scale value*. These three functions that Petr would have called "associating functions" satisfy the proposed new VIM definition of Direct Measurement for all Sciences. The term "Weak Associative Measurement" can now be replaced by Direct Measurement when the instrument transacts these three functions of *Transduction, Application of Instrument Scale,* and *Calibration to a Public Scale*. Petr's scientific modeling remains apt, and it holds for a strong definition (direct measurement), rather than a weak definition (indirect measurement).

6.1.2 (§56) The Latent Common Factor Prototype of Latent Variable Modeling

The LCF model will now be introduced as a special modeling tool in the frame of those previously presented aspects and methodological considerations depicted in Figures 6.1–6.3, and as a special case in the family of models with latent variables. First, let us go back to the example of identification of the latent variable *f* as "Verbal Ability" from the correlations in Table 3.2 of Section §34 in Chapter 3. Although we did not know the mechanism by which the hidden factor score of *f* had been calculated for each of the *N* subjects, we still may be satisfied that the manifest-latent correlations r_{jf} between the indicators x_j and the LCF *f* left only negligible residuals "unexplained" by their partial correlations with *f*. Let's try to characterize some specific features of the indicated approach:

- The objects of modeling, that is the indicators, consisted of the original manifest scores. These are the observed or measured numerical values x_j, and their accompanying matrix of correlation coefficients r_{jk}. In other words, the modeled objects were not proportions nor response probabilities.
- Consequently, the demand for completeness of explanation was focused on a specific type of local independence, that is, partial uncorrelatedness. This is linear correlation coefficient independence only, hence not general probabilistic independence.
- The theoretical concept of the "ability" to be identified was hypothesized as a scalable scientific variable *f*, hopefully fulfilling the representation-measurement conditions of an interval scale, measurable in terms of a latent *z*-score metric. Therefore, it is not any other type of measurement level, that is, not an ordinal or an unordered categorical (nominal) scientific concept. A nominal type of theoretical concept could be appropriately modeled in the case of a purely qualitative typology, perhaps by using a Latent Class Model rather than the LCF model.
- The syntactic/mathematical association functions standing formally for correspondence rules were conceptually chosen in the form of linear regression and expressed by a linear measurement model equation.

Considerations regarding the above-mentioned aspects should be included in an appropriate methodological description before and during the process of designing the research study or psychometric instrument.

It is a rather common habit, though not always appropriate, to work with correlation matrices rather than with covariance matrices in linear factor analysis, and to express not only the latent variable f but also the manifest variable indicators x in the z-standardized form. This usually seems reasonable in the light of "adding apples and oranges," especially if the observed scores of the indicators are in various measurement units and have different-sized standard deviations. Moreover, usually their measurement units have no realistically interpretable meaning for representation of the theoretical concept, that is the construct.[2] We will maintain this habit of working with the standardized z-scores in most cases.

From a general point of view, use of the LCF model means that semantic correspondence in its Theoretical → Empirical direction, formalized as a syntactic derivation (in Figure 6.2, on the right) gets an initial formal version, a mathematically approximate equation, namely a *general measurement model* (not yet fully specified):

$$x_j = F_j(f) + d_j$$

This works with certain *discrepancies* d_j as its errors in the equation. But the LCF model represents an even more specialized case of such an approximation, namely, it is a falsifiable *statistical model*, where each of $j = 1, ..., n$ functions $F_j(f) = \hat{x}_j$ is a *regression function* with the *errors of regression* d_j. Additionally, to characterize the LCF model fully in the frame of formalization of correspondence rules, all of its regression functions $F_j(f)$ are linear. As indicated above, if both manifest x_j, as well as latent f, variables are in z-standardized form, then the *linear regression* gets the form $\hat{x}_j = \beta_{j|f} f$, or $x_j = \beta_{j|f} f + d_j, j = 1, ..., n$, where $\hat{x}_j$ means a statistical regression estimate. The multiplier $\beta_{j|f}$ is the standardized regression coefficient, and the general unspecified "error in equation" d_j would take on its special form. In other words, d_j becomes the *error of regression estimate*, also called the *regression residual*.

Looking ahead to the viewpoint of Section §59 and its terminology for decomposing an observed score x_j into generic and unique parts by the equation $x_j = g_j + d_j$, it is clear that in the LCF model, $\hat{x}_j = \beta_{j|f} f$, f is a special case of the generic true part g_j; and d_j, the *unique* part, is a special variant of $e_{j|f}$, the residual of the regression of x_j on f. The residual d_j is a deviation of the observed score from its estimate produced by the would-like-to-be-known hypothetical level of "ability" represented by the LCF f in the position of an independent variable. In the LCF model, these regression beta coefficients, $\beta_{j|f}$, are usually denoted by a simpler symbolization (a_j) while sometimes being given more precise names in the usual jargon of this topic. As a standard, we will be referring to $a_j = \beta_{j|f}$ as the latent-common *factor loading*.[3]

Then the *LCF measurement model equation* becomes

$$\hat{x}_j = a_j f, \qquad \text{or} \qquad x_j = a_j f + e_{j|f}, \qquad j = 1, ..., n.$$

---

2 Sometimes it is reasonable to work with summing up the (0;1) raw scores of dichotomous items. This is to get a total score of a knowledge test or a didactic test, since a z-score transformation of the (0;1) scores is not adequate nor appropriate in most situations.

3 Among other names, the term "factor coefficients" is also quite common. But, sometimes this term might be misunderstood as "factor score coefficients." This term is used in SPSS where it indicates coefficients for constructing estimates of factor scores (as explained in Section §59). The term "loadings" comes from one of the founders of psychometrics, L.L. Thurstone, who had been formerly an electrical engineer working for T.A. Edison. After he switched to psychology and factor analysis, he still kept a metaphor of his multiple factors of primary mental abilities as electric batteries *loading* the individual tests. Another synonym for factor loadings is "factor saturations."

Thus, the LCF model represents a very simple kind of association function $F_j(.)$, which stands for the syntactic derivation in the overall modeling procedure for a particular type of correspondence rule. Modeling the scores of the jth manifest indicator is accomplished by *multiplying* the latent variable f by the *coefficient* a_j, and then adding the error component $e_{j\,|\,f}$.

6.1.3 (§57) How to Calculate Loadings and Residual Correlations in the Factoring Process

Back in Chapter 3, in the example with four verbal tests and the Verbal Ability factor (Table 3.2, Section §34), we introduced the process of obtaining a latent variable using factor analysis. We did not yet indicate how the artificial "explanatory" f had been found and simulated by a particular practical computational procedure at hand. The process might be explained as seeking and numerically constructing a latent score of f for each subject such that the mutual partial correlations between all the manifest x_j and x_k become zero after subtracting out the influence of f. This is the *demand for completeness of explanation*. It is known in general as the *requirement of local independence for latent variable models*. For the LCF model, it is particularly specified as the demonstration of *partial uncorrelatedness*.

As we know from Sections §32 and §33, the partial correlation $r_{12\,|\,f}$ of two variables x_1, x_2 can be zero if and only if we are able to find a third variable, here the factor f, whose correlations with x_1 and x_2 fulfill the *numerator check of local uncorrelatedness* $(r_{12} - r_{1f}r_{2f}) = 0$, which for factoring is expressed as $(r_{12} - a_1a_2) = 0$. In practice, this means that f leaves the "unexplained" residual of r_{12}, namely the $(r_{12} - a_1a_2)$ numerator of $r_{RES\ 12}$, close to 0. Hence, although we do not have the actual scores of the subjects on the latent variable f at our disposal, if we know only the correlation between the indicators r_{12} and the correlations of f with the indicators (r_{1f}, r_{2f}, that is, the loadings a_1 and a_2), we still may determine whether the variable f could do its explanatory job acceptably well. Therefore, *we need not know f, the latent scores* themselves, but *only their loadings* $r_{jf} = a_j$ to decide whether the latent variable f can explain the original intercorrelations.

6.1.3.1 Applying the Latent Common Factor Model

We now illustrate the computational procedures using the example of the four verbal tests from Table 3.2 of Chapter 3, Section §34, whose numerical steps are now repeated in Table 6.1. It is essentially the mathematics of the numerator checking for partial correlations. As a separate example, we take just one correlation coefficient $r_{12} = 0.50$ between test x_1 Verbal Fluency and test x_2 Boston Naming. We can see that its value of 0.50 was explained quite well by partialing out the latent f, which brought its residual close to zero, $r_{RES12} = 0.02$. This residual was calculated by using the *numerator check of local uncorrelatedness*, where a_1 and a_2 are the factor loadings for variables x_1 and x_2:

$$r_{RES_{jk}} = (r_{12} - a_1a_2) = 0.50 - 0.69 \times 0.70 \cong 0.50 - 0.48 = 0.02.$$

By analogy, the computations may be done for the other r_{jk} residuals. The correlations produced by the model are sometimes denoted as the "reproduced" correlations $r_{jk}^{+} = a_ja_k$. The difference

$$\left(r_{jk} - r_{jk}^{+}\right) = \left(r_{jk} - a_ja_k\right)$$

indicates how close these modeled reproductions r_{jk}^{+} are to the original correlations r_{jk}. The procedure is shown in Table 6.1 where the products a_ja_k for all pairs of indicators are placed in the reproduced correlation matrix $\left[r_{jk}^{+}\right] = \left[a_ja_k\right]$. This matrix is then subtracted from the original correlation matrix r_{jk} to produce the matrix of correlation residuals r_{RESml}. This displays the amount of

Table 6.1 The correlation matrix of four verbal tests (r_{jk}), the reproduced correlation matrix from a single maximum likelihood factor (r^{+}_{jk}), and the residual matrix ($r_{RES(ml)}$) reflecting the small amount of correlation left unexplained by an LCF of verbal competency.

	$[r_{jk}]$					$[r^{+}_{jk}] = (a_j)(a_k)$						$[r_{RES(ml)}] = [r_{jk}] - [r^{+}_{jk}]$				
	Original					**Reproduced**						**Residuals**				
	X_1	X_2	X_3	X_4		$a_j \backslash a_k$	0.69	0.69	0.67	0.71			X_1	X_2	X_3	X_4
X_1 Verbal Fluency		0.50	0.52	0.43	=	0.69		0.48	0.47	0.49	+	X_1		0.02	0.05	−0.06
X_2 Boston Naming	0.50		0.40	0.54		0.69	0.48		0.47	0.50		X_2	0.02		−0.07	0.04
X_3 Control Oral Word	0.52	0.40		0.50		0.67	0.47	0.47		0.48		X_3	0.05	−0.07		0.02
X_4 Word List Association	0.43	0.54	0.50			0.71	0.49	0.50	0.48			X_4	−0.06	0.04	0.02	

correlatedness remaining after subtracting out the amount of correlation explained by the product of the factor loadings for each a_j and a_k pair. As the residuals approach zero, we have evidence the factor explains the correlations.

How do we find these four loadings a_1, a_2, a_3, a_4? They are shown as the row and column marginal labels of matrix $\left[r_{jk}^{+}\right]$[4] in Table 6.1, where we also see from the residual matrix how well they identify the explained correlation. What is the rationale and the numerical procedure used to produce the vector of loadings [0.6916, 0.6962, 0.6764, and 0.7120]? If we had the correlation of just one of the pairs of variables given, $r_{12} = 0.50$, it would be easy to specify many pairs of numerical values a_1, a_2 that would produce the product 0.50. But we need more complex methods to identify four loadings that can jointly explain *all six correlations* of the nonredundant values in the lower triangle of the correlation matrix $[r_{jk}]$.

There are methods of numerical analysis that enable us to obtain a *computational optimization*. These algorithms are programmed and used in a large variety of optimization problems, not only in latent variable modeling. They usually use matrix algebra in a repeated, iterative process for improving the roughly guessed numerical starting values of a_j to reach a final optimal solution. The methods have names such as Newton–Raphson, Davidon–Fletcher–Powell, and Maximum Likelihood (ML). These numerical methods are incorporated into complex software packages, many of which are used in a variety of models for structural modeling, factoring, and related methods. Some of them are stand-alone commercial software programs, such as LISREL, EQS, and Mplus.[5] Some are procedures in general-purpose statistical software packages (SAS PROC CALIS, Stata's sem and gsem, SPSS Amos), and some are noncommercial programs (lavaan, openmx, and onyx) associated with the free statistical software package R.

6.1.3.2 Centroid Method of Factoring

Given the mathematical and computational complexity that has grown up around the factor analytic methods, it would seem that they have become a kind of black box, shrouded in mystery. However, going back in time, we find a method of factoring, a factoring algorithm, that is simple and easy to perform with only a calculator. It is on about the same conceptual difficulty level as calculating a variance or a correlation coefficient. It is called the *centroid method* and it was particularly popular over a half-century ago before electronic computers made other more effective models practical.

Notice from Table 6.2 how simple the centroid method is to calculate using the four steps spelled out in the left side of the table. First, one obtains the sums of the entries in each row of a target correlation matrix as shown in the right half of the table. One then calculates the sum of these sums as shown directly below the column vector of the obtained row sums. The square root of the sum of sums $\sqrt{16.24} = 4.03$ is then divided into each of the row sums to obtain the four-factor loadings a_j, as shown in the column on the far right: 3.09/4.03 = 0.77; 3.10/4.03 = 0.77; 3.15/4.03 = 0.78; 3.18/4.03 = 0.79; and 3.72/4.03 = 0.92.

---

4 Although the correlational entries in the three matrices, the original correlations $\left[r_{jk}\right]$, the reproduced correlations $\left[r_{jk}^{+}\right]$, and the residuals $[r_{RES(ml)}]$, are all shown in just two-place accuracy for simplicity, the loadings in the row and column marginal labels are given with four place accuracy. When the product between a_1 and a_2 is carried out to four place accuracy it comes out to $a_1a_2 = (0.6916)(0.6962) = 0.4815$, which when subtracted from the $r = 0.50$ correlation yields more precisely a residual correlation of $0.5000 - 0.4815 = 0.0185$, just a little less than 0.02.

5 Not all of these packages are available in 2025 or readily available. This book's computational guides on Factor Analysis, Classical Test Theory, and IRT use widely institutionally available Stata, SPSS, and MPlus packages.

Table 6.2 A simple algorithm for calculating the centroid factor loadings.

Step 1. Sum each row of the correlation matrix	$R =$	$X1$	$X2$	$X3$	$X4$	$X5$	Sums	a_j
Step 2. Sum the column of these sums and record it below the column (16.24)	$X1$	1.00	0.40	0.52	0.50	0.67	3.09	0.77
	$X2$	0.40	1.00	0.51	0.54	0.65	3.10	0.77
Step 3. Calculate the square root of this sum of sums and record it below (4.03)	$X3$	0.52	0.51	1.00	0.43	0.69	3.15	0.78
	$X4$	0.50	0.54	0.43	1.00	0.71	3.18	0.79
Step 4. To obtain the centroid factor loadings in column a_j, divide each of the five row sums by this square	$X5$	0.67	0.65	0.69	0.71	1.00	3.72	0.92
							Sum of sums = 16.24	
							$\sqrt{Sum\ of\ sums} = 4.03$	

The centroid method is the least complex of the factoring methods, perhaps also near the low end with regard to effectiveness, and is now only of historical interest. Lawley's (1943) ML method of factoring provides a stark contrast. It is precise, complex, and challenging to understand, but also particularly effective at identifying factor solutions that are locally independent. Interestingly, Lawley found the centroid method to be of use in the creation of the ML method. Commenting on the centroid method, Lawley and Maxwell (1962) state that "although a statistical assessment of the method and an examination of its sampling properties is virtually impossible because of a certain arbitrariness in its procedure, it often gives estimates of the loadings which are fairly close to ML estimates and which can be usefully employed when finding the latter." Lawley and Maxwell (1962) then describe the geometry of the centroid method in some detail and go on to outline the detailed and complex process of ML estimation as it builds upon but goes well beyond the simple centroid method.

6.1.3.3 Comparison of Three Models: The "Merely Historical Centroid Model," the PC Model, and the ML Model

The centroid method has been attributed to Thurstone, and indeed he was central to its development in the early part of the 20th century, but several others also were. As Lawley and Maxwell state, "the *centroid* or *simple summation* method is well described by numerous writers (cf. Burt, 1940; Thurstone, 1934; Thomson, 1951; Jowett, 1958)." Emmett (1949) also provides insight into the translation into and early historical development of ML factoring.[6] Of course, in today's world, ML is a major part of virtually every statistical data analysis software system. Table 6.3 compares the performance of three diverse methods of factoring, the centroid method (which is not used today except for pedagogical purposes), the principal-component method (which is based upon eigenvector/eigenvalue mathematics, but is not one of the LCF methods), and the ML method which is an

6 There exists a wide variety of factoring methods, most of which are based upon the eigenvectors/eigenvalues approach introduced by Karl Pearson (1901) in his initial paper on the principal component method. The method is reasonably understandable and ubiquitous. Brown, Hendrix, Hedges, & Smith (2012, pp. 153–163) give a "simplest case mathematical demonstration" of a complete principal component method factor analysis. Maximum likelihood is based upon a mathematical approach (Lawley and Maxwell, 1971) that diverges from, and is in important ways, more interesting than the many other methods based upon eigenvectors and eigenvalues. This is particularly true when the dataset is large, and a multivariate normal distribution can be assumed. Maximum likelihood seems to be unusually strong in identifying latent common factor models that fulfill the requirement of partial uncorrelatedness, the correlational (linear models) version of stochastic local independence.

Table 6.3 Comparison of centroid factors, principal component factors, and maximum likelihood factors with regard to factor loadings, original correlation matrix, reproduced correlation matrix, and residual correlation matrix.

Centroid method

Original $[r_{jk}]$						**Reproduced *R* Matrix** $[r^*_{jk}] = a_j a_k$						**Residual *R* Matrix** $[r_{RES(C)}] = [r_{jk}] - [r^*_{jk}]$					
	***X*1**	***X*2**	***X*3**	***X*4**	***X*5**		***X*1**	***X*2**	***X*3**	***X*4**	***X*5**		***X*1**	***X*2**	***X*3**	***X*4**	***X*5**
$X1$		0.400	0.520	0.500	0.670	$X1$		0.590	0.599	0.605	0.708	$X1$		−0.190	−0.079	−0.105	−0.038
$X2$	0.400		0.510	0.540	0.650	$X2$	0.590		0.601	0.607	0.710	$X2$	−0.190		−0.091	−0.067	−0.060
$X3$	0.520	0.510		0.430	0.690	$X3$	0.599	0.601		0.617	0.722	$X3$	−0.079	−0.091		−0.187	−0.032
$X4$	0.500	0.540	0.430		0.710	$X4$	0.605	0.607	0.617		0.728	$X4$	−0.105	−0.067	−0.187		−0.018
$X5$	0.670	0.650	0.690	0.710		$X5$	0.708	0.710	0.722	0.728		$X5$	−0.038	−0.060	−0.032	−0.018	

Principal-component method

$[r_{jk}]$ **Original**						$[r^*_{jk}] = a_j a_k$ **Reproduced *R* Matrix**						$[r_{RES(PC)}] = [r_{jk}] - [r^*_{jk}]$ **Residual *R* Matrix**					
	***X*1**	***X*2**	***X*3**	***X*4**	***X*5**		***X*1**	***X*2**	***X*3**	***X*4**	***X*5**		***X*1**	***X*2**	***X*3**	***X*4**	***X*5**
$X1$		0.400	0.520	0.500	0.670	$X1$		0.584	0.596	0.603	0.710	$X1$		−0.184	−0.076	−0.103	−0.040
$X2$	0.400		0.510	0.540	0.650	$X2$	0.584		0.597	0.605	0.712	$X2$	−0.184		−0.087	−0.065	−0.062
$X3$	0.520	0.510		0.430	0.690	$X3$	0.596	0.597		0.617	0.726	$X3$	−0.076	−0.087		−0.187	−0.036
$X4$	0.500	0.540	0.430		0.710	$X4$	0.603	0.605	0.617		0.735	$X4$	−0.103	−0.065	−0.187		−0.025
$X5$	0.670	0.650	0.690	0.710		$X5$	0.710	0.712	0.726	0.735		$X5$	−0.040	−0.062	−0.036	−0.025	

Table 6.3 (Continued)

Maximum Likelihood method																
$[r_{jk}]$						$[r^{+}_{jk}] = a_j a_k$						$[r_{RES(ml)}] = [r_{jk}] - [r^{+}_{jk}]$				
Original						Reproduced *R* Matrix						Residual *R* Matrix				
	***X*1**	***X*2**	***X*3**	***X*4**	***X*5**		**0.684**	**0.667**	**0.704**	**0.723**	**0.979**		***X*1**	***X*2**	***X*3**	***X*4**	***X*5**
*X*1		0.400	0.520	0.500	0.670	*X*1		0.456	0.481	0.495	0.670	*X*1		−0.056	0.039	0.005	0.000
*X*2	0.400		0.510	0.540	0.650	*X*2	0.456		0.469	0.482	0.653	*X*2	−0.056		0.041	0.058	−0.003
*X*3	0.520	0.510		0.430	0.690	*X*3	0.481	0.469		0.509	0.689	*X*3	0.039	0.041		−0.079	0.001
*X*4	0.500	0.540	0.430		0.710	*X*4	0.495	0.482	0.509		0.708	*X*4	0.005	0.058	−0.079		0.002
*X*5	0.670	0.650	0.690	0.710		*X*5	0.670	0.653	0.689	0.708		*X*5	0.000	−0.003	0.001	0.002	

LCF model, and one of the best. The ML method like many other methods uses a linear combination of variables to form factors, but the ML parameter estimates are derived in such a way as to be most likely to have resulted in the observed correlation matrix.

The reproduced correlation matrices in the middle column of Table 6.3 can be calculated one pair at a time, as $(a_j)(a_k)$, or more efficiently by postmultiplying the column vector of loadings a (used as column headers in the middle of Table 6.3) by the row vector a' (which is nothing but a in its transposed form as was demonstrated in Appendix 2 and as shown below for the ML data). The diagonal entries are greyed out of the reproduced correlations (r^+) matrix since they are not used in this process of finding residuals.

$$r_{jk}^{+} = aa' = \begin{bmatrix} 0.684 \\ 0.667 \\ 0.704 \\ 0.723 \\ 0.979 \end{bmatrix} = [0.684 \;\; 0.667 \;\; 0.704 \;\; 0.723 \;\; 0.979]$$

$$= \begin{bmatrix} 0.468 & 0.456 & 0.481 & 0.495 & 0.670 \\ 0.456 & 0.444 & 0.469 & 0.482 & 0.653 \\ 0.481 & 0.469 & 0.495 & 0.509 & 0.689 \\ 0.495 & 0.482 & 0.509 & 0.523 & 0.708 \\ 0.670 & 0.653 & 0.689 & 0.708 & 0.959 \end{bmatrix}$$

Notice the clear superiority of the ML method over the other two in terms of partial uncorrelatedness as reflected in the values of the residual correlation matrices $r_{RES_{jk}}$. Among the ten values in the upper triangular part of the ML residual matrix (or for that matter, in its mirror image lower triangular part), only one residual is larger in absolute value than 0.06 ($r_{RES_{34}} = -0.08$). On the other hand, the principal component r_{RESjk} matrix has five out of ten larger in absolute value than 0.06. The centroid r_{RESjk} matrix has six out of ten larger than 0.06. The average r_{RESjk} value for the ML analysis is 0.028, which is less than one-third of each of the other average values, 0.086 for principal component analysis (PCA), and 0.087 for centroid method analysis.

6.1.3.4 Fundamental Factor Theorem

Table 6.3 provides a comparison of these three factor analytic methods all applied to the same input correlation matrix. The ML factoring algorithm is one of the best methods for obtaining partial uncorrelatedness in fitting an LCF model. Mathematically, this "completeness of the explanation" can also be viewed from an opposite perspective. Fulfillment of the requirement of partial uncorrelatedness need not be considered only a desired target, but can also be mathematically and logically viewed as a *consequence* of certain intrinsic properties of the data at hand. Namely, it is possible to derive mathematically that IF:

i) the common factor f is uncorrelated with the unique parts d_j of the indicators and
ii) the unique parts d_j, d_k of various indicators x_j, x_k are mutually uncorrelated

THEN it will follow for all the pairs of indicators that $(r_{jk} - a_j a_k) = 0$. In other words, it follows that when these conditions are met, the residual correlation matrix for the dataset *will necessarily have* local independence.

This is known as the *Fundamental Factor Theorem*. It is also called Thurstone's Factor Theorem to recognize L.L. Thurstone who derived it from the above-mentioned two assumptions. The two assumptions may be recognized to have the character of the LCF model axioms: (i) $r(f, d_j) = 0$ for all j, (ii) $r(d_j, d_k) = 0$ for all $j \neq k$. More intuitively, if the unique parts are supposed to be truly unique, they must not correlate with the uniqueness of other indicators or with the common parts. In this perspective, it is interesting that local independence as embodied in partial uncorrelatedness follows as a consequence of the implicit mathematical background of the model.

Thus, the two axioms dealing with the uncorrelatedness of common and unique factors become more than just the usual favorite assumptions made by statisticians. In the case of the LCF model, the consequence derived from its axioms makes the model intrinsically suitable for explanatory purposes since it also implies the principle of partial uncorrelatedness, which might at first appear to be an "accidental" consequence of the model. On the other hand, for a successful application of the model, the two axioms should have some substantive, theoretical as well as empirical, interpretation.

6.1.4 (§58) Maximize the Fit of the Model and Minimize the Discrepancy Index to Achieve Local Independence

As observed in the previous section, the average residuals value $r_{RES_{jk}}$ for the ML analysis in Table 6.3 is quite small at 0.028, less than one-third as large as the averages of 0.086 and 0.087 for PCA and centroid analysis, respectively. In other words, it comes closer to meeting the criterion of local independence. The vague formulation "on average" suggests that an overall characteristic could be based on a *fit function* that averages the residuals, whose value should be minimized. One of the possible usable functions is the fit function based on the *sum of squares of the residuals*, essentially a variation measure, written as $SS(r_{RES})$, and referred to as $Q_{RES} = SS(r_{RES})$[7].

We can compare $SS(r_{RES}) = 0.016$ for the ML analysis to $SS(r_{RES}) = 0.104$ for the PCA and $SS(r_{RES}) = 0.108$ for the centroid factor analysis. For the ideal case, a matrix of Q_{RES} values of zero would mean a *perfect fit* where all of the original correlations are completely explained and our hypothetical structural model would fit the principle of partial uncorrelatedness perfectly. The empirical value of Q_{RES} may be interpreted as a "distance" of the pattern in the correlation matrix from the ideal solution, that is, as an amount of discrepancy from our factoring criterion. Therefore, this kind of fit function is sometimes called a *discrepancy function*, or a "penalty function," or also a *loss function* (the last term coming from economic applications).

6.1.4.1 Least Squares Fitting and Model Fit Assessment

The procedures for finding optimal solutions are constructed mathematically so that they are decreasing a fit function step-by-step until its possible minimum can be found. The optimal solution for the loadings in Table 6.3 were based on such an explicit numerical search for those values of a_j that yield the smallest or so-called least value of $SS(r_{RES})$. Therefore, Q_{RES} is called the *Least Squares* fit function, abbreviated as the *LS-function*. The set of finally computed loadings a_j, found at the point where Q_{RES} had reached its minimum, is referred to as a *Least Squares solution*.

Of course, the values of the fit function for correlation matrices of different dimensions, that is, for datasets with different numbers of indicators, are not directly comparable. Hence, the above-mentioned intuitive concept of an "average" of correlation residuals is now coming back. What may serve this purpose is literally the mean squared residual. A still more practical variant is its

7 We can also refer to *sum of squares residual*, $SS(r_{RES})$, as Q_{RES}, that is, *"quantity" for residuals.*

square root, which could be interpreted intuitively as an "average spread of unexplained correlation." This is called the *Root Mean Squared Residual (RMSR)*. It is calculated from the residual matrices in the rightmost position of each of the three analyses shown in Table 6.3. In more modern usage, this is abbreviated as *RMR*, [8] where the "*M*" is understood to stand for "mean square." It is symbolized and calculated as $RMR = \sqrt{Mean(r_{RES}^2)} = \sqrt{1/M^* Q_{RES}}$ where Q_{RES} stands for SS_{RES}, that is, "sum of the squared residuals." and $M^* = n^2 - n = 5^2 - 5 = 20$ is the number of coefficients in the 5×5 residuals matrix, $r_{RESjk}, j \neq k$. For example, the RMR value for the ML analysis at the bottom of Table 6.3 is $RMR = \sqrt{1/20 \times 0.0318} = 0.040$. A rule of thumb may be to think of a good fit when the residual correlations are "explained" below a particular value of *RMR* such as 0.05. In addition, it is good to check whether there are not some individual high $|r_{RES}|$ values hidden within the overall "average" *RMR*. Sometimes the *maximum* $|r_{RES}|_{\max}$ value is also reported. In Table 6.3, the $|r_{RES}|_{\max} = 0.08$ for the ML analysis, somewhat lower than the $|r_{RES}|_{\max}$ values of 0.19 each for the other two analyses.

6.1.4.2 Relative Model Fit

Another aspect of the degree of fit for different data deals with the overall magnitude of the original manifest correlations to be modeled. If the range of r_{jk} in Table 6.3 is from 0.40 to 0.54, maybe "explaining" them out to the average level $RMR = 0.035$ could be relatively acceptable. But, what if the original correlations were quite low? Or how would one compare that *RMR* value to the case where the original correlations were around 0.90? Therefore, another index is based on the relative comparison of the fit function of residual correlations $[r_{RES\,jk}]$ to the value of the fit function applied to the original correlations $[r_{jk}]$.

This would mean computing *SS* of the original correlation matrix $\boldsymbol{R} = [r_{jk}]$, in which case it is usual to use the notation $Q_0 = SS[r_{jk}]$, and designate the "fit" function of "no-model" as the *base line model*, that is, when there is actually no restriction, no model was fitted at all. On the other hand, we have the correlation matrix *generated by the model*, $\boldsymbol{R}_{\mathrm{m}}$ and its sum of squares $Q_{\mathrm{m}} = SS[r_{\mathrm{m}\,jk}]$, hence $Q_{RES} = Q_0 - Q_{\mathrm{m}}$. Then the relative proportion of the "unexplained" correlations is given by the ratio $SS[r_{RES}]/SS[r_{jk}]$, that is, by $Q_{RES}/Q_0 = (Q_0 - Q_{\mathrm{m}})/Q_0$ which may be assumed to be comparable across different data matrices.

For instance, in Table 6.3, we compute the $Q_0 = SS[r_{jk}] = 6.537$ for the original correlation matrix analyzed (which is equivalent for all three factoring methods in that table). Calculating the "relative proportion of the unexplained correlations" for the ML analysis, we then get $Q_{RES}/Q_0 = 0.032/6.537 \cong 0.0049$. It means that of the original overall "amount of intercorrelation," the proportion of correlation left as residual, is only 0.0049 (about 0.5%)[9] while 99.5% has been "explained." The latter variant is often used, though not in percents but as a proportional ratio, whose value approaching 1.00 indicates proximity to the perfect fit. In older writings, this index was generally known as Bentler's Δ (delta), introduced by the two authors, Bentler and Bonett, 1980. Recently, it is known as one of the relative *Goodness of Fit Indices* (GFI) usually printed out by most SEM computer programs, often presented under the name Normed Fit Index, *NFI*:

$$GFI = NFI = (Q_0 - Q_{RES})/Q_0 = 1 - Q_{RES}/Q_0$$

8 *RMR* values (formerly called *RMSR values*) are not comparable across datasets unless they are standardized. Our use of *RMR* here actually is standardized, however, since we are applying the RMR formula to correlation matrices as shown in Table 6.3.

9 0.0049 (49 ten thousandths), compared to 0.0319 (319 ten thousandths) for principal component method, and 0.0329 (329 ten thousandths) for the centroid method.

This has a clear intuitive interpretation. In this example, the value of the Least Squares index $GFI = NFI = 1 - \frac{Q_{RES}}{Q_0} = 1 - 0.0049 = 0.9951$ indicates an "excellent fit" by generally accepted standards.

Although the index was introduced here just for the LS function, it also has analogous versions for *other types of fit functions*. Therefore, the variant presented above can be specified as GFI_{LS}. Almost all other fit indices are standardized to yield values from 0 to 1 following the *GFI*'s intuitive logic of a kind of *proportion explained* by the model divided by the overall magnitude. Let us recall that one of the possible interpretations of the correlation coefficient r_{xy} follows this common sense of a proportion, namely the computed covariance divided by the maximum covariance, or also in its r_{xy}^2 version a ratio of explained variance divided by original variance. In this sense, r_{xy} may also be understood as an index of the same nature since it expresses the fit of a model, namely the model of a regression line fitted to the scatter plot of *(x, y)* data points.

6.1.4.3 Unidimensionality and Content Homogeneity

The level of fit in the above-mentioned example supports the formal hypothesis of latent unidimensionality of the four manifest variables in a sense of the linear LCF model. Consequently, it supports our theory that Verbal Ability might be a content-homogeneous minor subdomain of a macro-behavior as it is diagnosed by clinical neuropsychological tests for Alzheimer's disease. If the fit were too bad, one would have to reject the hypothesis that the set of indicators is unidimensional—measuring jointly only one latent variable—regardless of the fact that the former expertise had suggested that the indicators should have been content homogeneous. It could be the case that the indicators measure more than just one theoretical concept and/or that the set contains some indicators that are influenced too much by a mixture of other unknown attributes, conditions, etc. (not necessarily simply due to unreliability). Nevertheless, it is important that the mathematical form of the model selected be conceptually well-based and specified. Namely, it could be the case that the theory is not quite wrong in principle from the substantive-semantic point of view, but its specification in its syntactic formulation of a chosen mathematical model could be inappropriate. This depends on the type of the model's latent scale, the linear vs. nonlinear association function, and other properties of the model as mentioned in this chapter. For example, instead of the LCF linear model, a more appropriate model could be an IRT model or so on. In this respect, it seems strange how often the LCF or PCA models are applied without any explicit rationale or a minimum content-elaborated theory about the investigated domain.

6.1.4.4 Weighted Least Squares (WLS) Fitting

The Least Squares fit function is directly focused on the target given by the desired principle of local-partial uncorrelatedness as the requirement of completeness of explanation. However, some other fit functions do not have this principle as an explicit primary target. Nevertheless, they also fulfill the principle, having it implicitly contained in their rationale as well as in the numerical procedures. From the list of currently used fit functions, let us name two here:

- Generalized least squares, known also as Weighted least squares, *GLS* or *WLS fit function*,
- Maximum likelihood, *ML fit function*.

The *GLS/WLS function* is similar to the LS function, but it gives different weights to individual residual correlations before summing up their "unexplained" squares. WLS methods generally give larger weights w_j to the indicators x_j whose common overall loading is larger. By intuition, it means

that the indicators with an overall *higher factor validity* will also have their residuals *better "explained"*, which is desirable due to their greater importance (relative to the other indicators). In the single-factor situation presented up to now in this text, such weights given to different indicators x_j, x_k are chosen to be

$$w_j = 1/\sqrt{1-a_j^2}, w_k = 1/\sqrt{1-a_k^2}.$$

Thus, the larger the magnitude of the factor validity (loading) a_j, the larger its importance weight w_j. From a numeric point of view, the WLS method may be seen as a generalization of the ordinary, unweighted, LS method with only the exception that it is *not* applied to residual correlations in the former matrix $[r_{RES\ jk}]$ but to the weighted residuals $[w_j r_{RES\ jk} w_k]$. It means the WLS method looks for the loadings a_j that, instead of minimizing $Q_{RES} = SSq[r_{RES\ jk}]$, minimize

$$Q_{\boldsymbol{W}ES} = SSq\left[w_j r_{RES\ jk} w_k\right].$$

Another interpretation could be that $a_j^2 = r_{jf}^2$ is a coefficient of determination of f by a single separate indicator x_j, while $\left(1-a_j^2\right)$ represents nondetermination of the (unit) variance of f. So, its square root is the standard error s_e of estimation of f. Hence, the smaller the standard error, the larger the weight $w_j = 1/s_e$. Recall from Sections §30 and §46 that $\frac{1}{s_e^2} = 1/\left(1-a_j^2\right)$ is the Fisher information statistic. Therefore, it may be said that in the WLS method, the weights w_j maximize the information which the indicators x_j offer about f. In addition, as we will see in what follows (Section §60), the nondetermination is usually denoted by $\left(1-a_j^2\right) = u_j^2$ and called the "uniqueness" of the indicator j. Then we may also write $w_j = 1/u_j$ and give the weight an alternative interpretation: in a common homogeneous cluster, the indicators that have larger "uniqueness", with respect to other indicators, also have less "importance", with respect to having their mutual relationships "explained."

6.1.4.5 Maximum Likelihood (ML) Fitting

It is also important to note that the WLS estimates of loadings a_j have additional desirable mathematical properties. In particular, these WLS estimates converge to the desired solution obtainable by the ML method. The motivation for obtaining the *ML solution* is based on the framework of inferential statistics dealing with the relationship between a sample and population (see Section §13). The ML method searches for the values of loadings a_j whose likelihood is maximum. Loosely speaking, this method's goal is to offer the maximally probable solution with respect to the population. However, it is interesting that it can be shown mathematically that for a large sample size N, the two differently driven conceptions, WLS and ML, converge to the same results. Asymptotically, the WLS minimization of $Q_{WRES} = SSq[w_j r_{RES\ jk} w_k]$ also maximizes the ML fit function offering the "most likely estimates" of the loadings. At this point, allow us to underscore the interesting agreement between the two different conceptions: In the WLS solution, it seemed reasonable to give more importance to the indicators with higher factor ("construct") validity—and this principle is found to be the best one from the point of view of reaching the "most probable" solution in the statistical sense of the ML fit function.

The minimization of a discrepancy of an LCF solution corresponds to the attempt to approach the ideal of "explaining out" the indicators' intercorrelations as far as possible (with the given, hypothetically prescribed, LCF structural model). In so doing, we fulfill the requirement of partial uncorrelatedness, the correlational version of stochastic local independence. In the case of the single-factor model, the requirement could be graphically pictured by a path diagram similar to the one shown in Figure 3.5b (Section §36) of Chapter 3, but now illustrating the *latent* "explanatory" variable f in Figure 6.4.

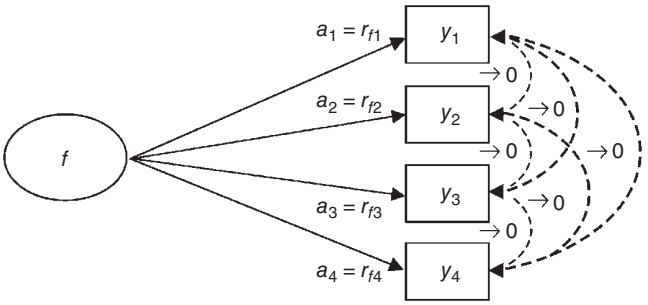


Figure 6.4 Path diagram of a well-fitted single factor LCF model approaching partial uncorrelatedness, that is, with the residual correlations close to zero (→0).

6.2 Factor Scores

The first half of this chapter has presented several ways to identify an LCF that is capable of explaining the correlations among the set of indicators. Once the common factor has been defined, the next order of business is to estimate scores; that is, factor scores of the respondents on that common factor.

6.2.1 (§59) Estimating Each Subject's Latent Score *f* to Measure a Trait or Characteristic

Imagine the point of view of somebody who would be standing behind the curtain, so to speak, and would know the hidden scores of subjects on *f*, say their levels of "ability." If that somebody were an advocate of the causal or propensity interpretations, those levels of *f* seem to be in the position of a source that produces the displayed performances coded by scores on observable indicators x_j. On the other hand, a spectator in the auditorium may just speculate about how to estimate or approximate the latent scores of *f* in individual subjects given knowledge of the manifest variables. What might be the numerical procedure which is used by a computer to find values of *f* that bring the indicators' intercorrelations close to zero after *f* has been "partialed out"? Given the observed scores of the indicators and the common-factor validities (loadings), how would one construct estimates of the latent scores? Does only one optimal construction of the latent scores exist, or may there be several different sets of such approximate latent score constructions that fit equally well?

An answer to these questions is still more complicated as it is related to a wider problem of so-called fundamental *factor indeterminacy*.[10] Nevertheless, even now we may say that a spectator's knowledge of the manifest scores and the common-factor validities (loadings) of a successfully fitting LCF model is not enough to recover the exact, determined knowledge that would have been at the disposal of the fictitious observer behind the curtain. We may just construct certain approximations. Those are sometimes very crude, as the simple sum and/or an intuitively based "formula" of generic scores of Section §58. In the best case, we may construct an *estimate* based on rational or even optimal assumptions, but in any case, such an estimate is not necessarily identical to the "true" latent score. Moreover, every time there is more than one set of estimates that fulfill the principle of partial uncorrelatedness (local independence) equally well and may thereby function as the latent "explanatory" variable, it means that factor scores of individual subjects cannot be "determined" or

10 Several later sections in the original Blahuš manuscript dealt with factor score indeterminacy. They were not included in this book. See comments in the Summary at the end of Chapter 6 on the importance of this issue and the greater utility of IRT for obtaining better values of person scores than are obtainable with LCF factor scores.

"established," they can only be estimated (however, the optimal formulas for estimating them might be recognized as a kind of "formula score" in a broader sense). However, in the context of latent variable modeling, any such optimal formula for constructing estimates represents an estimate, and not the true f itself—since the true score is indeed latent.

6.2.1.1 Estimating Factor Scores

Just as there are various criteria of optimality in estimating loadings, there are also varying approaches to the estimation of latent factor scores. For simplicity, we will present only a method based on the WLS principle, the so-called *Bartlett's WLS method*, shown in Table 6.4.

The Bartlett (1938) method has several mathematical advantages, as its solution converges to the ML method and the errors of its estimates are uncorrelated with the "true" unknown f. Its rationale may be intuitively interpreted in terms similar to the estimation of loadings in Section §58, in which larger weights b_j for the indicators were assigned to the subtests that had larger squared loadings a_j^2, that is, larger *coefficients of determination*. However, the weighting for the calculation of factor score coefficients corresponds to the larger *coefficients of nondetermination* calculated from the loadings, $1 - a_j^2 = u_j^2$ as shown in the equations in the upper right of Table 6.4.

The three-step process for obtaining factor score coefficients begins with the factor loadings, a_j, which are shown in the top center of Table 6.4. We use Eq. (6.3) from the set of three equations shown in the upper right of Table 6.4 to calculate q from the factor loadings a_j of the four subscales:

$$q = {}^{1}\Big/ \sum \frac{\left(1 - a_j^2\right)}{a_j^2} = {}^{1}\Big/ \left[\frac{\left(1-0.70^2\right)}{0.70^2} + \frac{\left(1-0.71^2\right)}{0.71^2} + \frac{\left(1-0.69^2\right)}{0.69^2} + \frac{\left(1-0.72^2\right)}{0.72^2}\right] = \frac{1}{4.05} \cong 0.25$$

Each of the factor score coefficients, b_j, for each of the indicators can now be calculated from its loading, a_j, combined with the q value just calculated. Then, using Eq. (6.2), the first coefficient, b_1, becomes

$$b_1 = \left[\frac{a_1}{1 - a_1^2}\right] q = \left[\frac{0.70}{1 - 0.70^2}\right] \times 0.25 = 0.34 \approx 0.35$$

(where the value 0.35 is the more precise one. It was calculated by a computer, hence it is not influenced by rounding errors). This formula may then be applied to the other three indicators $b_2 = 0.36$, $b_3 = 0.33$, $b_{42} = 0.38$, etc., as shown in the bottom of Table 6.4.

These four factor score coefficients are then used according to Eq. (6.1) $\left(\hat{f}_{BLS} = \sum b_j x_j\right)$ to demonstrate in the lower left of Table 6.4 the calculation of estimates of the factor scores for the first two persons. The estimation of the factor scores as values on the latent interval scale of $\hat{f}$ is one of the practical outcomes of an application of LCF modeling in its role as a "device" for scaling latent variables. In the sense of a theoretical and substantively based choice of the conceptually appropriate LCF model, we are able "to *estimate* the level of Verbal Ability," as shown in the example of Table 6.4. For instance, the estimated level of the first individual is 1.19 z-points (standard deviations) above the hypothetical mean of that theoretical concept, Verbal Ability, in the sample. The estimates $\hat{f}$ differ from the true unknown f by regression errors $e_f = f - \hat{f}$, and we have an ordinary regression equation for f: $f = \hat{f} + e_f$, where $r_{\hat{f}e_f} = 0$.

This indicates that errors are uncorrelated with $\hat{f}$ as specified by the definition of regression. Using the correlation matrix $\boldsymbol{R}$ of indicators $x_1, \ldots, x_n$, and their factor loadings $a_1, \ldots, a_n$, it is possible

Table 6.4 Computation of the *estimates* $\dot{f} = \hat{f}_{BLS}$ of the factor scores by the Bartlett Least Squares "construction formula" by first creating factor score coefficients, b_j. The coefficients are then used to create a linear combination of *Z*-scores of the indicators using matrix multiplication. The process is explained in the commentary accompanying this table.

Manifest variables x_j representing the indicators for a test of Verbal Ability					**Factor Loadings**	**Latent Common Factor *f* modeling the theoretical concept of "Verbal Ability"**
	Subscales	Raw score units	Mean	SD	a_j	
x_1	Verbal Fluency	[no. of words/min]	11.1	5.4	0.70	$f_{\text{BLS}} = \sum b_j x_j$ *(Eq. 6.1)*
x_2	Boston Naming	[no. of namings]	11.9	3.0	0.71	$b_j = a_j/(1 - a_j^2)q$ *(Eq. 6.2)*
x_3	Control Oral Word	[no. of associations]	23.6	11.8	0.69	
x_4	Word List	[no. of words]	4.3	2.0	0.72	$q = 1/\sum(1 - a_j^2)/a_j^2$ *(Eq. 6.3)*

Raw Scores *for the first five persons*					**Standardized Scores *for the first five persons***			
Person	x_1	x_2	x_3	x_4	z_1	z_2	z_3	z_4
1	23	15	19	5	2.20	1.03	−0.39	0.35
2	12	12	9	5	0.17	0.03	−1.24	0.35
3	8	14	19	4	−0.57	0.70	−0.39	−0.15
4	19	13	28	3	1.46	0.37	0.37	−0.65
5	13	14	25	4	0.35	0.70	0.12	−0.15

Estimates of the latent *f* (Bartlet factor scores)

Person	f =	b_1z_1 +	b_2z_2 +	b_3z_3 +	b_4z_4
1	1.15 =	0.35 × 2.20 +	0.36 × 1.03 +	0.33 × (−0.39) +	0.38 × 0.35
2	−0.20 =	0.35 × 0.17 +	0.36 × 0.03 + 0.33 × (−1.24) +		0.35 × 2.21
3	−0.13 =	...			
4	0.52 =	...			
5	0.36 =	...			

Multiple correlation for estimating *f* from the indicators x_1 to x_4: $R_{f|1,2,3,4} = 0.896$.
Standard error of the regression estimate of *f* by f_{BLS}: $S_{ef} = S_{f|1,2,3,4} = 0.44$.

to compute the multiple correlation $R_{f|\,1\ldots n}$, and $s_{e_f} = s_{f|1\ldots n}$, the standard error of the regression estimates of f as reported at the bottom of Table 6.4 according to the methods demonstrated in Appendix 5, Multiple Regression.

6.2.1.2 Factor Scores and Latent Values

Now, we are almost in the position of somebody who is standing behind the curtain, "the person who knows"—but only almost, due to the imprecision of our knowledge about f, since the standard error of its latent z-score is about half of an SD in this example ($s_{f|\,1,\,2,\,3,\,4} = 0.44$). So, knowing at least the estimates of the latent values f for the subjects, we may hope to generate the manifestations of f in the performance scores x in the four verbal tests based on our knowledge of the level of "Verbal Ability." The scores can be reproduced by the very LCF model that was introduced in Section §56, namely by the regression equation $\hat{x}_j = a_j f$. Of course, we recognize that this is just a variant of the z-score regression equation of Appendix 5, Multiple Regression, and in Section §30 of Chapter 3, $z_{y\ est} = r_{xy} z_x = \beta_{y|x} z_x$. The only specialization of that regression equation is that the independent variable f is now latent. Since our information on f is contaminated by the errors e_f, we may try to substitute the unprecise estimates $\hat{f}$ for f into the regression equation $\hat{x}_j = a_j f$ creating thus a "doubled estimation":

$$\hat{\hat{x}}_j = a_j \hat{f}.$$

For example, in Table 6.4, in test 1, subject number 1 has the z-standardized score $x_{11} = 2.23$. What would be his/her estimate from the would-like-to-be-known f_1 score substituted by the "unprecise" $\hat{f} = 1.19$? Since the factor loading for indicator 1 is $a_1 = 0.70$, the equation produces $\hat{\hat{x}}_{11} = 0.70 \times 1.19 = 0.83$ as an "estimate" of the observed z-value $x_{11} = 2.23$.

How is it possible that the LCF model, which was fitting the correlations quite well, is generating such a "bad" estimate? There are two reasons:

i) The first one is that the independent variable $\hat{f}$ had already been distorted by its deviations from the true f before entering into the equation. This means that the first source in the sequence is the uncertainty of estimation of f from the manifest variables. Therefore, the independent variable here is not the true f, but only its estimate.
ii) The other one—and this is conceptually more important—is that the model focuses implicitly on exhausting only the information obtained from the indicators which is common to them, and may be used for the *construction of the common* latent factor that *"explains" the correlations*, primarily. This built-in tendency has the two following consequences.

With regard to the original manifest data, i.e. the scores x, the LCF model is not focused on reproducing them completely but only on their *generic part* g_j, which is called the *common part* in the special case of the LCF model. Only these common parts g_j are then made "responsible" for reproducing the correlations $[r_{jk}]$. Thus, as already mentioned in Section §58, the LCF model *is <u>not</u> focused on reproducing the variances* $s_j^2 = 1$ in the diagonal of the original correlation matrix (the unities, due to z-standardization). Driven by the "explanatory" principle of local-partial uncorrelatedness, it searches for such a latent variable f that might be supposed to be jointly measured by the indicators—which assumption would account for why the indicators are correlated. Thus, the LCF modeling process does not seek a variable, which would "explain" the original data, the scores, but only their correlations. This important feature distinguishes LCF modeling from PCA description, the analysis aimed at the decomposition of variances into principal components.

6.2.1.3 Generic True Scores and the Latent Common Factor Model

Let's assume for a while that the "unprecise" value $\hat{f}$ may be used as the would-like-to-be known latent "true" score of individual number 1, and so assume that $f_1 = 1.19$ can be the correct latent score. Then in the sense of Section §59, and of the LCF equation, that $\hat{\hat{x}}_{11}$ would stand for the value $g_{11} = 0.84$ and the decomposition of that individual's score x_{11} on the indicator x_1 would become a specialization of

$x_j = g_j + d_j$	namely	$x_{11} = g_{11} + d_{11}$
		$2.23 = 0.84 + d_{11}$
		$2.23 = 0.84 + 1.39.$

It means that the model would take only the amount of 0.84 from that subject's score 2.23 to use it for modeling—i.e. reproducing—the correlations in the matrix $[r_{jk}]$. The rest of the subject's score, $d_{11} = 1.39$ is the unique part, which does not help the model generate the correlations. While the individual's score is quite high on this test, being 2.23 SD above the mean, such a performance is rather "unique" in the context of all the indicators as regards the substantial contribution to the identification of their common latent factor—only a small magnitude (0.84) of the whole magnitude of the score (2.23) may be used. Of course, these numbers and the calculation above represent just a kind of virtual reality. Such a fantasy could be brought even farther along if we imagine that we knew which part of the unique $d_{11} = 1.40$ is an error of observation ε_{11} in that subject's score, and could play a game as if we might compute the true τ_{11} and specific δ_{11} components.

The reader should keep in mind that the above consideration using the examples with would-like-to-be known scores of f in the "doubled backward re-estimation of x" are speculations of just an illustrative meaning. It is a little bit of a strange fact that by statistical theory, we may get some information about the properties and consequences of the estimation of f, but the substitution of the concrete values of the estimates $\hat{f}$ in the place of the true f produces necessarily different results. Simply put, even the best estimate is not the same as the original.

6.2.2 (§60) Communality, the Variance Shared Among Indicators

In the sense of additive decomposition of the manifest score into its parts, as shown above, a corresponding decomposition of the variances can also be done. It extends the variance decomposition in Classical Test Theory, which was used to define reliability (Section §48). Since we consider the LCF model to be a case of linear regression, namely the regression equation $\hat{x}_j = a_j f$ or $x_j = a_j f + d_j$ with $a_j = \beta_{j|f}$, we have d_j uncorrelated with f by the definition of regression. Then, together with CTT and regression theory (Sections §48 and §30), we need just a couple of additional, but reasonably acceptable, weak assumptions to derive that the variance of a manifest variable may be decomposed into just a simple sum. Let's start with the variances corresponding to the generic part $g_j = \hat{x}_j = a_j f$ and its unique distortion d_j. Then, correspondingly to the decomposition of a manifest score x_j presented in Section §59,

$$x_j = g_j + d_j, \text{resp}\,. x_j = a_j f + d_j,$$

we get a similar decomposition of its variance $s^2_{x_j}$ (equal to 1 as x_j is in z-scores):

$$s^2_{x_j} = s^2_{g_j} + s^2_{d_j} = s^2_{(a_j f)} + s^2_{d_j} \equiv h^2_j + u^2_j = 1.$$

By the convention of LCF modeling, the following special terms are in use:

- *communality*, $h_j^2 = s^2_{(a_j f)}$ for the variance of the generic—common part of x_j and
- *uniqueness*, $u_j^2 = s^2_{d_j}$ for the variance of the unique part.

In a sense, uniqueness, the unique variance of an indicator, may be rather interestingly interpreted in a negative way (Harman, 1960, p. 14). 'It, ... indicates the extent to which the common factors *fail* to account for the total unit variance of the variable.

6.2.2.1 Communality, Specificity, and Unreliability

Recall from Section §59 that the unique distortion d_j of the indicator´s score is composed of two additive parts, $d_j = \delta_j + \varepsilon_j$, the reliable *specific* part δ_j, and the *unreliable* part ε_j, the *error* due to unreliability of observation/measurement (in the sense of the specific true score, cf. Section §48). Accordingly, the overall uniqueness u_j^2 of the jth indicator is the part of its observed variance, $s^2_{x_j}$, which is unique to the given concrete observation of the indicator's scores. It is composed of two additional parts of variance:

$$u_j^2 = u^2_{\delta_j} + u^2_{\varepsilon_j}, \text{i.e.} \{\text{Uniqueness}\} = \{\text{Specificity}\} + \{\text{Unreliability}\},$$

- *specificity*—the variance of δ_j, the still reliable part of indicator's variance but solely specific to this jth indicator and
- *unreliability*—the variance of ε_j, i.e. the variance of observation errors due to the indicator's specific unreliability (its individual, separate unreliability).

Due to z-standardization, $s^2_{x_j} = 1$, the proportion of the error variance $u^2_{\varepsilon_j}$ to the total observed variance follows the CT-theory definition of specific unreliability of a separate test/indicator:

$$u^2_{\varepsilon_j} \equiv Unrel(x_j),$$

as it was defined Section §45, Section §48 by $Unrel(x_j) = u^2_{\varepsilon_j}/s^2_{x_j}$. Similarly, from the viewpoint of the CTT model, $x_j = \tau_j + \varepsilon_j$, the proportion of true variance $s^2_{\tau_j}$ in the total unit variance $s^2_{x_j} = 1$, is the reliability $Rel(x) = s^2_{\tau_j}/s^2_{x_j}$ itself. Recall from Section §59 that $\tau_j = g_j + \delta_j$, whose variance may be decomposed correspondingly:

$$s^2_{\tau_j} = Rel(x_j) = h_j^2 + u^2_{\delta_j}$$

Thus, we get the well-known assertion about the LCF model that "Reliability consists of communality plus specificity." The calculated concrete values of the communalities $h_j{}^2$ are obtained as part of a standard solution of a fitted LCF model in the associated computer output. Thus, if an indicator's individual specific reliability is already known (say by the test–retest method), then it is also easy to compute the indicator's specificity by

$$u^2_{\delta_j} = Rel(x_j) - h_j^2 \equiv s^2_{\tau_j} - h_j^2$$
$$\{\text{Specificity}\} = \{\text{Reliability}\} - \{\text{Communality}\}.$$

It is possible to say that the specific variance of an indicator identifies and distinguishes its score from the other indicators of the same common latent factor, it is its idiosyncratic property, and it determines its identity with respect to the other tests. An expression for obtaining communalities and uniquenesses of individual indicators is especially easy in the case of just *one* LCF:

$$h_j^2 = a_j^2.$$

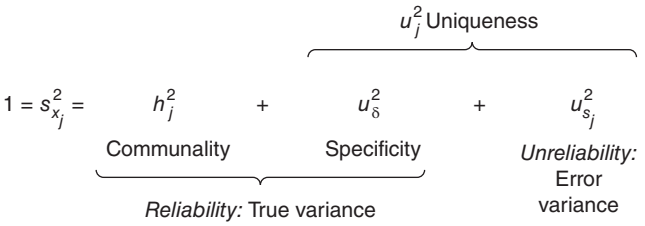


Figure 6.5 The decomposition of the manifest variance of an indicator.

In the single-factor model, this may be derived in a very simple way (though a little nontraditional). Finally, an overall scheme of the variance components can be presented as shown in Figure 6.5.

The decomposition of a score was already shown earlier in the example with the Verbal Fluency test in Section §59. Let's repeat the illustrative hypothetical decomposition of a score of subject number 1 in test 1:

$$x_{11} = g_{11} + d_{11}, \text{i.e. } 2.23 = 0.84 + 1.39.$$

If we imagine that we knew such a partition in every subject $i = 1, ..., N$, we could calculate the communality h_1^2 of indicator 1 by the ordinary formula for the variance of the values g_{1i}. But in reality, we are able to do it just by application of the LCF model using our knowledge of the factor loadings for one latent factor, by using $a_1 = 0.70$ in the formula above:

$$h_1^2 = a_1^2 = 0.70^2 = 0.49$$

We can likewise calculate the uniqueness as:

$$u_1^2 = 1 - h_1^2 = 0.51.$$

This can be interpreted in a way that only 49% of the overall diagnostic variability of V-Fluency test x_1 as an indicator is contributing to a "joint effort" of all the indicators to identify their common latent variable f ("Verbal Ability") in terms of this model of "associative" measurement. The rest, $u_1^2 = 0.51$, is variance unique to the application of this subtest. This uniqueness is a mixture of the subtest's specific character, and the errors due to the unreliability of the procedure of its use: $u_1^2 = u_{\delta_1}^2 + u_{\varepsilon_1}^2$. Unfortunately there is no accessible information from standardization studies on the V-Fluency test about its reliability $Rel(x_1)$. Therefore, the error variance $u_{\varepsilon_1}^2 = Unrel(x_1)$ cannot be subtracted from the uniqueness u_1^2 to compute the specificity. This will be illustrated in an example with different data in the next section.

6.2.2.2 Communality and the Squared Multiple Correlation (SMC) with the Other Indicators

Let us stress that the notion of the adjective "common" in the "*common* part" of a score and/or the "*common* variance", also known as communality should be understood properly. The *common* part is the "generic part," i.e. that part of an indicator's score which is "explained" by the LCF (in the sense of local-partial uncorrelatedness of this indicator with the other ones). On the other hand, an equivalent to *communality* is the "generic variance" of the latent factor scores of the subjects. It is necessary to underscore that this communality is "common" in the sense that the indicator shares it with the other indicators in their "common contribution" to estimating their common latent factor f. Or, in the opposite direction, it is that part of the indicator's variance that can be "explained" by

the LCF. Hence, this is part of the variance that a test, as an indicator of *f*, has in common with other indicators, due to an intermediating role of the factor, or "through the factor *f*."

However, a quite different meaning of the word "common" might be used in the sense of such a part of the test variance which would be estimable by other tests in the sense of the so-called SMC, the Squared Multiple Correlation coefficient. *SMC* is the multiple *coefficient of determination*, that is, $R^2_{1|234}$. In the sense of Section §30, it offers information about the proportion of the variance of a subtest x_1 that is estimable by multiple regression on the rest of the tests in the battery. In a sense, it represents a common proportion of a test's variance, here of the variance of x_1, that the test has "directly in common with other tests", even without any idea that the subtests were conceptualized to be indicators of any particular latent factor *f*. Hence, the SMC represents a sort of internal validity of the test, possibly showing redundancy of the battery. These two different meanings of the words "common variance" might create confusion since the proportion of variance estimable by the SMC (i.e. R^2) is proved to be lower than the communality or maximally equal to it. In the "before computers" early times of factor analysis, the SMCs of separate indicators x_j were applied as an a-priori "SMC-guess" of the communalities. More recently, it is well known that SMC is a *lower bound of communality*, that is, an indicator's communality cannot be smaller than its SMC, that is, its squared multiple correlation with all the other indicators in the battery (Rencher, 2002, p. 422).

$$h^2_j \geq R^2_{j|1,2,\ldots>j<\ldots n},$$

Intuitively, this fact may be understood to be due to a "vacuum cleaner effect" of the common latent factor to extract, especially "all the common information." For example, we might suspect that the battery of the four verbal tests is somewhat redundant, and even that one of the subtests, perhaps the V-Fluency test x_1 could be abandoned. As shown in Section §30, we may assess composite validity, that is, the internal validity of the rest of the three tests with respect to x_1, by multiple regression and compute the multiple correlation $R_{1|2,3,4} = 0.62$ (the actual value in that research study). Then we find that its corresponding SMC or coefficient of determination is $R^2_{1|2,3,4} = 0.62^2 = 0.38$, which satisfies the inequality: $0.43 \geq 0.38$.

On the other hand, this inequality, together with the value of the reliability of a specific subtest in a battery, can help to roughly approximate the possible subtest's specificity by its upper bound even without application of the LCF model:

$$\{\text{Specificity}\} \geq \{\text{Reliability}\} - \{\text{SMC}\}.$$

Some additional inequalities suggest themselves, for instance {Reliability}≥ {Communality}, in which the lower bound of a specific separate subtest's reliability is usually so much lower that it becomes of little practical use—we would like to hope that reliability of the V-Fluency test is much higher than 0.48.

6.2.3 (§61) An Example Regression Estimate of *f* and the Decomposition of Indicator Variance

In several past sections, we were dealing with $Rel(x_j)$, the reliability of a separate diagnostic procedure, meaning the specific reliability of an isolated test or subtest x_j along with the standard error of measurement s_{ε_j} of its specific true score τ_j. By the assumptions of CT Theory, an approximation of the specific reliability can best be done under the assumptions of parallel forms by the test–retest method, with $Rel(x_j)$ approximated by the test–retest correlation $r_{jj'}$ (see Sections §47 and §48). Now, assume that we have a content-homogeneous cluster of indicators that were fitted by one latent factor, and that test–retest repetitions of any of the indicators x_1, ..., x_n can be assumed to produce

their parallel counterparts so that we could have parallel pairs of indicators. This will be exemplified with empirical data below.

6.2.3.1 The Tapping Test

In his introductory course in test theory, Professor Blahuš used data obtained by administering several simple tests to the students. A favorite was a set of tapping tests carried out with just a pencil and paper. Table 6.5 includes obtained data from $N = 12$ graduate students in each of $n = 4$ different tapping subtests. Each of the subtests was three seconds long as measured on a watch, and controlled by verbal instructions to "Go" and "Stop." (Note that this is also an example of several potential sources of unreliability in a testing session.) The task was to produce as many dots as possible in the three seconds by use of the left hand only (subtest x_1), by the right hand only (subtest x_2), by both hands alternatingly with two pencils (subtest x_3), and by both hands simultaneously (subtest x_4).

The structural hypothesis was that these four indicators were content-homogeneous, measuring a theoretical concept—"ability," which was named by the students' Motor-Frequency ability, that is, the "MF ability test." Since an "ability" should be a scalable theoretical-conceptual variable, and a linear form of the association-measurement function seemed to be provisionally acceptable, the LCF model had been chosen to test several properties of the battery. Unidimensionality was assessed by fitting a latent common f, to assess construct validity of indicators in terms of their latent factor validities (loadings) a_j, and to model the level of "MF ability" by computing estimates of the latent scores for f in the twelve individuals who were tested. We used the SPSS program to estimate the loadings using the LS method and used the BLS method (by Bartlett, 1938) for estimating f scores as shown in Section §59. In addition, from the data in Table 6.5, the multiple correlation for estimating f from the four indicators was computed. We also calculated its square, the coefficient of determination, and the *standard error of the regression estimate*:

$$R_{f|1\ldots n} = 0.968, \qquad R^2_{f|1\ldots n} = 0.937, \text{and } s_{e_f} = \sqrt{1-0.937} = \sqrt{0.063} = 0.25.$$

This means that the latent factor score can be estimated approximately with standard error ± 0.25 z-scores, a quarter of an SD. Under the assumption of a normal distribution, the typical *range of errors* when estimating f can be approximated by using the rule of ± 3 sigmas, or the popular 0.95 probability rule of ± 2 sigmas / $\pm\, 2s_{e_f}$.

6.2.3.2 Test–Retest Analysis

After ten minutes, the retest session was administered to these twelve students. Petr obtained analogous results on that retest occasion, which are presented in what follows, as given in the left part of Table 6.7. These were used for computation of test–retest correlations $r_{jj'}$ which could be considered to approximate the subtest reliabilities $Rel(x_j)$ given that the retests are parallel forms of their first-occasion counterparts. Just for the illustrative purpose of the example, let's assume parallelism, and accept the $r_{jj'}$ in Table 6.6 to be reliabilities.

Following the CT Theory formulas in Section §48, we can calculate the corresponding standard errors of the separate subtests. For instance, for subtest x_1, we have

$$s_{e_1} = s_{x_1}\sqrt{1-Rel(x_1)} \approx s_{x_1}\sqrt{1-r_{jj'}} = s_{x_1}\sqrt{1-0.955} = s_{x_1}\sqrt{0.045} \cong s_{x_1} \times 0.21,$$

meaning 0.21 in z-scores, i.e. about one-fifth of an SD. But there is a problem here because the SDs of exactly parallel forms should be equal, $s_1 = s_{1'}$, which is not the case, since by comparing Tables 6.5 and 6.7, we find $s_1 = 3.54$ while $s_{1'} = 4.38$. Of course, in practical situations, we never find exactly equal SDs, but this difference seems to be big enough to contradict the assumption

Table 6.5 One-factor LCF model of the four tapping tests on the first test occasion. Estimates of the latent scores *f* are modeled as measures of motor-frequency ability, by using unweighted least squares factoring.

	Raw Scores				**Standardized *z*-scores**						**Correlations [r_{jk}]**				**Factor Validities**
Person	1	2	3	4	x_1	x_2	x_3	x_4	f		x_1	x_2	x_3	x_4	a_j
1	12	15	28	30	−1.13	−1.31	−0.63	−0.65	−1.19	x_1	1				0.898
2	22	23	48	37	1.69	0.86	2.04	0.46	1.46	x_2	0.842	1			0.923
3	12	15	20	29	−1.13	−1.31	−1.70	−0.80	−1.45	x_3	0.765	0.772	1		0.851
4	15	21	35	42	−0.28	0.32	0.30	1.25	0.28	x_4	0.666	0.703	0.664	1	0.760
5	17	21	27	26	0.28	0.32	−0.77	−1.28	−0.10						
6	19	23	37	41	0.85	0.86	0.57	1.09	0.93						
7	17	24	39	37	0.28	1.13	0.84	0.46	0.85						
3	21	25	40	45	1.41	1.40	0.97	1.73	1.54	Residual Correlations [r_{RES}]					
9	12	18	33	26	−1.13	−0.50	0.03	−1.28	−0.76		x_1	x_2	x_3		
10	17	17	29	33	0.28	−0.77	−0.50	−0.17	−0.39	x_2	0.013				
11	16	21	29	33	0.00	0.32	−0.50	−0.17	0.01	x_3	0.001	−0.014			
12	12	15	28	30	−1.13	−1.31	−0.63	−0.65	−1.19	x_4	−0.016	0.001	0.016		
	Raw score statistics									***FIT:***					
j =	1		2		3		4			GFI = 0.99					
Mean:	16.00		19.83		32.75		34.08			Q_{LS} = 0.0009					
SD:	3.54		3.69		7.49		6.32			RMSR = 0.01					
										Max $\|r_{RES}\|$ =0.016					

Table 6.6 The test–retest approximations of the indicator-specific reliabilities for the four indicators.

j/j'	1	2	3	4
$r_{jj'}$	0.955	0.938	0.979	0.943

about parallelism of the repeated test. Nevertheless, just for illustrative purposes, we will proceed as if the variances were almost equal and use their mean as an approximation:

$$s_{x_1} \approx \sqrt{\frac{s_1^2 + s_{1'}^2}{2}} = \sqrt{(3.54^2 + 4.38^2)/2} = 3.98 \cong 4[\text{taps}], \text{hence}, s_{\varepsilon_1} \cong 4 \times 0.21 = 0.84\ [\text{taps}].$$

Thus, the roughly approximated standard error of subtest x_1 is smaller than 1 [tap] in the raw score units of the tapping test, which would seem acceptable "precision."

Continuing on, the indicator-specific reliabilities now allow us to compute the specific variance—the *specificity*—of each indicator (which was not possible with the Verbal Ability tests, as we did not know their individual specific reliabilities). Thus, the decomposition of the manifest variance indicated in the preceding Section §60 in general can be now expressed numerically using Table 6.7 for the indicator x_1, in Figure 6.6.

6.2.4 (§62) Generic Reliability and the "Model–Remodel" Assessment

The semantic-content validity of indicators together with joint—generic—*semantic content homogeneity of the whole test cluster* is usually verified by the use of an appropriate latent variable model, either a confirmatory LCF model, or Item Response Theory (IRT) models such as Latent Ogive, or so on. McDonald (1970, 1989), but especially McDonald (1982, p. 214), stresses the importance of using this latent modeling approach:

> A Spearman factor analysis can therefore be made the basis for generic reliability ..., that is, of generalizability in the sense of Cronbach et al. (1972) (see McDonald, 1978a). If ... the factor loadings ... are equal, the tests are essentially tau-equivalent, ... and the common factor differs only trivially from a specific true score. The factor analysis then supplies a basis for the assessment of reliability in the classical sense of measurement error.

Then the (linear) mathematical *unidimensionality* of the cluster is accepted as a modeling confirmation of the hypothesized semantic-content homogeneity. It means we then face the problem of *generic reliability*, i.e. of the *measurement error of the theoretic-conceptual variable* as it was modeled by the latent variable, say the LCF.

A conjecture may arise that after having fitted a single-factor LCF model on the first testing occasion to the correlation matrix $[r_{jk}]$ with latent factor f, why couldn't we try to fit one factor again, this time to the retest occasion correlation matrix $\left[r_{j'k'}\right]$ to obtain a new latent factor f'? Then, one may imagine that an analog of "test–retest" could be "model–remodel" approximation of *generic reliability Rel*(f) by $r_{ff'}$.

A comparison of Tables 6.5 and 6.7 indicates that the set of subtests was unidimensional on the first occasion as well as on the second (retest) occasion with a very good fit of the LCF model in each case. In the same way as in Table 6.5, the "remodeled" scores f' are computed from the retest data in

Table 6.7 Retest tapping scores and re-modeling of the latent scores f', to be correlated with the latent scores f in Table 6.5 to obtain a reliability estimate for the Blahuš motor-frequency ability test.

	Raw Scores				Standardized *z*-scores						Correlations [r_{jk}]				Factor Validities
Person	**1′**	**2′**	**3′**	**4′**	x'_1	x'_2	x'_3	x'_4	f'		x'_1	x'_2	x'_3	x'_4	a_j
1	10	17	30	29	−1.50	−0.77	−0.40	−0.99	−1.11	x'_1	1				0.867
2	23	23	50	37	1.47	0.58	2.00	0.41	1.39	x'_2	0.764	1			0.880
3	12	14	19	35	−1.05	−1.44	−1.72	0.06	−1.46	x'_3	0.726	0.715	1		0.825
4	15	24	35	42	−0.36	0.80	0.20	1.28	0.42	x'_4	0.560	0.597	0.547	1	0.662
5	18	21	29	26	0.32	0.13	−0.52	−1.51	−0.16						
6	22	27	37	40	1.24	1.48	0.44	0.93	1.31						
7	18	24	38	36	0.32	0.80	0.56	0.23	0.64						
8	22	25	45	45	1.24	1.03	1.40	1.80	1.50		Residual Correlations [r_{RES}]				
9	11	18	32	28	−1.28	−0.54	−0.16	−0.16	−0.89		x'_1	x'_2	x'_3		
10	16	16	29	32	−0.13	−0.99	−0.52	−0.47	−0.66	x'_2	0.001				
11	18	22	28	33	0.32	0.35	−0.64	−0.29	.06	x'_3	0.011	−0.012			
12	14	14	28	33	−0.59	−1.44	−0.64	−0.29	−1.03	x'_4	−0.015	0.014	0.001		
	Raw score statistics									***FIT:***					
$j =$	1′	2′	3′	4′						GFI = 0.99					
Mean:	16.58	20.42	33.33	34.67						$Q_{\text{LS}} = 0.0006$					
SD:	4.38	4.46	8.34	5.73						RMSR = 0.008					
										Max $\lvert r_{RES}\rvert = 0.015$					

Generic reliability of the latent score approximated by the model–remodel correlation: $r_{ff'} = 0.986$.

Communality: $h_1^2 = a_1^2 = 0.898^2 \cong 0.792$

Uniqueness: $u_1^2 = 1 - h_1^2 = 1 - 0.792 = 0.208$

Unreliability: $1 - Rel_1 = 1 - 0.955 = 0.045$

Specificity = Reliability – Communality = 0.955 – 0.792 – 0.163

= Uniqueness – Unreliability = 0.208 – 0.045 – 0.163

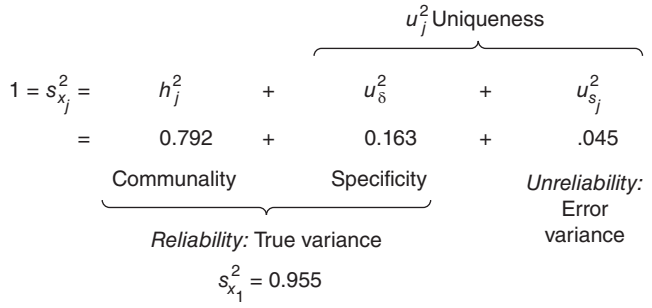


Figure 6.6 A numerical example of the decomposition of the manifest variance of a particular indicator, Indicator 1, which has a reliability of $r_{11'}$ of 0.955 as shown in Table 6.6. The reliability for the whole test, $r_{ff'}$, is 0.986 as given at the bottom of Table 6.7.

Table 6.7, and the "model–remodel" correlation $r_{ff'} = 0.986$ is presented as an approximation of the generic reliability of "associative" measurement of "Motor Frequency ability."

Since the value of f is estimated on a hypothetical scale in latent z-score units (SDs), the precision of measurement of f can be expressed by the *latent standard error* s_{ε_f}. Using $r_{ff'} = 0.986$ as an approximation of *Rel* (f), we find

$$Unrel(f) = s_{\varepsilon_f}^2 = 1 - 0.986 = 0.014, \text{and } s_{\varepsilon_f} \cong 0.12.$$

This means that measurement of the theoretic-conceptual "MF ability", obtained by modeling in terms of the LCF model, can establish the z-score levels of that "ability" in different individuals with a standard error 0.12 of the latent SD of f. Again, by the popular 0.95 probability rule of ± 2 sigmas / $\pm 2s_{\varepsilon_f}$, we could approximate the *range of errors* due to the *generic unreliability* of f to be $\pm 2s_{\varepsilon_f} = \pm 0.24$.

This approach to approximating generic reliability has a certain special feature. Namely, as pointed out in Section §59, the modeled (as well as the remodeled) latent scores are constructed to be *only estimates* of the unknown f, therefore they should actually have been symbolized by $\hat{f}$, and $\hat{f}'$. In this example, they were constructed by the Bartlett (1938) method. Hence, they are specified as $\hat{f}_{BLS}$, and $\hat{f}'_{BLS}$. As was already mentioned, these BLS estimates of true latent scores f have certain desirable mathematical properties especially if they are applied to the loadings a_j computed by the WLS method in Section §58. As compared to the other kinds of estimators, $\hat{f}$, one of the desired properties of $\hat{f}_{BLS}$ is that only Bartlett (1938) estimates ensure that the errors e_f of estimation of the unknown true latent f be uncorrelated with them. Thus, by defining $e_f = \hat{f}_{BLS} - f$, the property becomes $r_{e_f f} = 0$, similar to its counterpart of uncorrelatedness between the error and true score, $r_{\varepsilon_j \tau} = 0$, in the CT theory of reliability (Sections §46–§48). Another convenient property of the BLS estimator is its conditional unbiasedness, meaning that the $\hat{f}_{BLS}$ estimates are not systematically shifted as compared to the unknown true latent scores f. Hence, if the level of "ability" stays constant over

occasions, these two properties support the idea of using the "model–remodel" correlation coefficient $r_{\hat{f}\hat{f}'}$ between the two estimations of f as an approximation of $Rel(f)$, the generic reliability of f, on the same basis as the test–retest correlation approximates the specific reliability of a separate indicator: $Rel(f) = 0.986$, $s_e = 0.12$, with the latter being measured in the latent standard deviations.[11] Looking at the problem from the point of view of measurement, we cannot expect to reach $Rel(f) = 1$, and it seems to be clear that it is not possible to "establish" the latent scores of f with "certainty" or to measure them exactly.

6.3 Summary

Chapter 6 takes us farther into the latent factor model and its methods and mathematics than in prior chapters. Petr does so soundly in main content and structure, but not in his choice of the type of measurement to feature in sharing his extensive latent factor model knowledge. In Figure 6.3 we used the term *Weak Associative Measurement,* but then fully replaced it with the three instrument processes used in the Mari, Wilson, and Maul (2021) definition of Direct Measurement. It fits Petr's aspirations and conceptual model with few changes and offers great benefits and clarity.

The material in the sections of Part 6.1 that we did keep, concerning various approaches to estimating factor loadings, is important and you should be aware of the assumptions and various slants these approaches take. But probably more important is your proficiency with the software package or packages you will use as a professional in building the factor structures for future studies. Hopefully, the data you or others in a development project will have collected to support domain models will validate or lead to improvements in those models. Domain models of persons in particular, and also models of tasks and evidence, guide analyses of indicators that characterize skills in each domain. You may wish to use software packages such as those employed in Computational Guides 2 and 3 to run factor analyses and IRT analyses.

The summary comment at the end of Chapter 5 still holds. It is often easier to use factor analysis, preferably confirmatory, to identify the factor structure of your domain, and then develop measurement instruments in the light of how many factors you find are needed. One correlated factor may enable you to control for unwanted variance in studies. Another might shed light on the domain processes that are involved in a property, a skill, or a proficiency.

The section explaining the Bartlett (1938) method for obtaining factor scores for individuals provides additional perspective on why IRT has been increasingly chosen for latent variable work since Petr's time. Given factor structure information, it is convenient and powerful to select the indicators from the factor studies that can be put into an instrument with a new sample and estimate the item parameters using IRT. It is a nonlinear model closely associated with probabilities and provides more accuracy, utility, and stability. You not only get the item difficulties, but the person proficiencies as well, and on the same logit scale. There is no need to figure out the best method to use for factor scores, nor to worry about factor invariance. Petr mentions factor invariance and has materials that explain it that were not included in this book because of space limits. Hoping that factors will be invariant from sample to sample, and with different sets of indicators, is often disappointing when the data are analyzed. However, once an instrument has been developed and put into use, its factor structure can hold up over time even when new tasks or items are needed and added.

11 These advantages can often be realized in actual research when you have a large enough sample size for the assumptions of the Bartlett (1938) model, but will typically not apply to the small N examples we use in the chapters to demonstrate the models in a more simplified form.

6.4 Study Questions

A. Twenty-two Discussion Questions

Q1 Explain how semantic correspondence rules relate to syntactic correspondence rules. (Section §55)

Q2 How is the partial correlation formula related to testing for local independence? (Section §57)

Q3 Explain how the *numerator check for local independence* in partial correlation is related to the calculation of residual correlations for the latent common factor model. (Section §57)

Q4 How does local independence follow logically from whether or not the two axioms of the Fundamental Factor Theorem are valid for a given dataset? (Section §57)

Q5 What was Leon L. Thurstone's major contribution to understanding the rationale of the latent common factor model? (Section §57)

Q6 Explain the meaning of *Root Mean Squared Residual (RMSR, now called RMR)* as a fit function. It is explained in Section §58 under the subsection heading "Least squares fitting and model fit assessment," using data from Table 6.3. Describe verbally how it is calculated from the formula, $RMR = \sqrt{Mean\left(r_{RES}^2\right)} = \sqrt{1/M^* Q_{RES}}$, where Q_{RES} stands for SS_{RES}, that is, "sum of the squared residuals," and how it is used. *Now go to the follow-up question EQ6 in the seven Quantitative Exercise questions.* (Section §58)

Q7 The next subsection within Section §58, titled "Relative model fit," follows up on *DQ6* by providing a formula that contextualizes the *RMR* by connecting it to the root mean square of the *original correlation matrix* from which it is derived. Thus, it becomes the *relative proportion of unexplained correlations* as given by the ratio $SS[r_{RES}]/SS[r_{jk}]$, which may be assumed to be comparable across different data matrices. The formula for that ratio is $Q_{RES}/Q_0 = (Q_0 - Q_m)/Q_0$. Explain the process of calculating this *relative proportion of unexplained correlations* from the matrix of residual correlations. *Go to the follow-up question EQ7 in the seven Quantitative Exercise questions.*

Q8 What is the new name for Bentler's Δ (delta)? How is this index calculated from Q_{RES}/Q_0, that is, from the *relative proportion of unexplained correlations*? How is it interpreted? How is it now used and under what name? (Section §58) *Go to the follow-up question EQ8 in the seven Quantitative Exercise questions.*

Q9 How are weighted least squares (WLS) fitting of factor loadings similar to maximum likelihood (ML) fitting? Would you expect them to give similar results? (Section §58) *Go to the follow-up question EQ9 in the seven Quantitative Exercise questions.*

Q10 How is partial uncorrelatedness related to stochastic local independence? Which is a stronger position? Why? (Section §58)

Q11 How does the Bartlett (1938) WLS rationale for assigning weights to factor scores compare to or differ from the WLS rationale for assigning weights to factor loadings? (Section §59)

Q12 What are the three steps for obtaining factor score coefficients by the Bartlett (1938) method? (Section §59) *Go to the follow-up question EQ9 in the seven Quantitative Exercise questions.*

Q13 Describe the Bartlett (1938) Least Squares construction formula for factor scores. (Section §59)

Q14 What did you learn from Petr Blahuš's tapping test? Why did he do both a first tapping test and also a second iteration of the tapping test? (Section §61)

Q15 How was Professor Blahuš able to calculate the reliability approximation for the whole test from the test–retest data? (Section §61) *Go to the follow-up question EQ15 in the seven Quantitative Exercise questions.*

Q16 How was he able to calculate the reliability approximations shown in Table 6.6 for each of the four indicators from the test–retest data? (Section §61) *Go to the follow-up question EQ16 in the seven Quantitative Exercise questions.*

Q17 Why are the indicator-specific reliability approximations necessary for the calculation of the specific variance, that is the *specificity,* of each indicator as shown in Figure 6.6? (Section §61)

Q18 At the beginning of Section §62, what does McDonald say would be a necessary condition for Dr. Blahuš' test and its retest to be "tau-equivalent"?

Q19 As explained at the end of Section §62, why was the Bartlett (1938) method of calculating factor scores a particularly good way to utilize the assumption of tau-equivalence in estimating reliability?

Q20 As explained in the last two paragraphs of the chapter Summary, enumerate some of the advantages of using IRT to obtain factor scores after first identifying the factor structure with a factor analysis, "preferably confirmatory"? Why is confirmatory factor analysis preferred in this case?

Q21 Why does IRT's nonlinear model "provide more accuracy, utility, and stability" in the estimation of factor scores? You may wish to discuss this in the light of your answer to question *DQ19*.

Q22 Comment on the relative value of each mentioned advantage of IRT in the process of obtaining factor scores. How does it help with the issue of "factor invariance"?

B. Seven Quantitative Exercise Questions

EQ1 Draw upon your response to *DQ6*. Using the same data from the maximum likelihood analysis shown in Figure 6.3, follow your explanation of how the *RMR* is calculated and use the formula $RMR = \sqrt{Mean\left(r_{RES}^2\right)} = \sqrt{1/M^* Q_{RES}}$ to obtain *RMR* for the maximum likelihood

analysis of this dataset. Do you obtain the same answer as that given in the "Least squares fitting and model fit assessment" subsection, *RMR* = 0.040? Notice that the only thing needed as input to obtain this RMR index was the matrix of correlation residuals $[r_{RES(ml)}]$. You first squared each of the 20 off-diagonal entries in the residual matrix, then followed the formula to obtain the mean of those squared values, and finally took the square root of that mean to obtain the *root mean square residual*, *RMR*.

You are now ready to practice obtaining the *root mean square residual* of an additional dataset. Go to the second computational demonstration, *CD2*, which is a maximum likelihood analysis of the first tapping test (shown in Table 6.5 in the first part of Section §61). In the SPSS output, under the heading "Reproduced Correlations," you will find two matrices. The first is the matrix of reproduced correlations $[r^{+}_{jk}]$ and the second is the residual correlation matrix $[r_{RES(ml)}]$. Use the *RMR* formula to calculate the *root mean square residual* for this first tapping test. How does it compare to the *RMR* you found in this question a few minutes ago for the maximum likelihood analysis of the 5 × 5 matrix from Table 6.3?

EQ2 This is a computational follow-up on your response to discussion question *DQ7* on "Relative model fit." Please return to the computational demonstration of *RMR* you did in *EQ6* from the maximum likelihood analysis in *CD2* of Table 6.5 tapping test data. Following the steps you summarized in *DQ7* calculate the *relative proportion of unexplained correlations,* that is, Q_{RES}/Q_0.

EQ3 Drawing upon and following your verbal response to *DQ8,* transform your *relative proportion of unexplained correlations*, that is, the calculation of Q_{RES}/Q_0 in *EQ7*, into a *proportional ratio, whose value approaching 1.00 indicates proximity to the perfect fit*. This index, formerly known as Bentler's Δ (delta) has been recently known as one of the relative "Goodness of Fit Indices" (GFI) often printed out on SEM computer programs under the name "Normed Fit Index" (NFI). The formula for its calculation is: $GFI = NFI = (Q_0 - Q_{RES})/Q_0$. Calculate the NFI index on the tapping data from *EQ7* using this formula.[12]

EQ4 To examine how similar the results from and WLS factoring are to maximum likelihood factoring first go to the "Computational demonstrations using *SPSS* commands" and copy the *SPSS* syntax in *CD2*, the factor analysis of the first tapping data. Go to *SPSS*, enter this code in the *SPSS* syntax window and run it. You will see that the output will agree with the output in *CD2* but there will be more tables in the full *SPSS* output.[13] As you can see in the third to last line of syntax, "/extraction=ml," this analysis was done with maximum likelihood factoring.

Now add one more line to your syntax, "list", and re-run the analysis. *CD3*, the factor analysis for the second tapping data has this same additional line of syntax, and you will notice that it produces a listing of the original variables, and their z-score counterparts, along with the estimated factor score for each person. You are now ready to create one more run of this SPSS factoring code on the first tapping data, but this time running it using the WLS

12 Notice that the NFI index was 0.9951, an "excellent fit," for the demonstration data in the chapter, which was calculated on the maximum likelihood data from Table 6.3. When you have completed the calculation of the NFI index for the tapping data, you will find that it will be 0.9996 which is very high. What do you think is the reason for this very strong fit?

13 Not all of the output from a particular demonstration run in *SPSS* is included in the "Eight Computational Demonstrations" of the Study Questions. We selectively included various output components according to our purpose.

or generalized least squares (GLS)[14] method of factoring rather than maximum likelihood, by using this command for extraction: "/extraction=gls." Now carefully compare the results of the maximum likelihood factoring with the results of the WLS factoring. Compare the factor loadings for the two methods, and the factor scores for the two methods (created by the "list" command line). You can also use the original correlation matrices from your output in combination with the correlation residual matrices to create the "Normed Fit Index" (NFI) as you learned in *EQ8*. How do these two extraction methods compare in their NFI goodness of fit?

EQ5 The mathematical basis of the Bartlett (1938) least squares "construction formula" is not overly complex and may be worth familiarizing yourself with it as explained in the text accompanying Table 6.4. First look over Eq. (6.3) for obtaining q and then calculate it for yourself either on a spreadsheet or a hand calculator. Next, use q in conjunction with the factor loadings a_j in Eq. (6.2) for calculating the b_j values (the factor score coefficients), one for each of the indicators. Finally, use Eq. (6.1) to obtain estimated factor scores either using scalar algebra to calculate one person's factor score at a time, or using matrix multiplication (as shown in Appendix 3) to post-multiply the $N \times 4$ matrix of Z scores (N persons by 4 indicators) by the 4×1 vector of factor score coefficients.

EQ6 How did Professor Blahuš calculate the reliability estimate for the whole test score using the first and second tapping test data? He ran a correlation coefficient between the f' column of Table 6.5, the first tapping test, and the f' column of Table 6.7, the second tapping test. Using either a spreadsheet or a calculator run that correlation coefficient and compare your answer to the $r_{ff'} = 0.986$ reliability estimate that is recorded at the bottom of Table 6.7.

EQ7 How did he calculate the reliability estimates for each of the four indicators? He ran a correlation coefficient for each of the indicator pairs across the two tables: $r_{11'}$ for the first indicator; $r_{22'}$ for the second indicator; $r_{33'}$ for the third indicator; and $r_{44'}$ for the fourth indicator. Calculate these four correlations and compare them to the "indicator-specific reliability" estimates in Table 6.6.

C. Eight Computational Demonstrations

Computational demonstrations using *SPSS* commands

CD1 *Factor analysis of the correlation matrix from the analysis of the four verbal tests of Table 6.1 using SPSS*. Summary data for four verbal tests are presented in Table 6.1 consisting of a correlation matrix of four verbal tests. In the SPSS syntax below, we have also added the means and standard deviations for these four variables. As shown in the "Applying the latent common factor model" subsection of Section §57, this data can then be analyzed to obtain the communalities, factor loadings, and a variety of other statistics. We illustrate the calculations here first with SPSS syntax commands and later do this in Stata. In SPSS, we first use the matrix data command to read and label the variables, indicating that the data consist of (a full) correlation matrix of the variables, along with their means and standard deviations, for a sample of 200 individuals. The factor command is then used to request factor analysis of

14 In SPSS, the weighted least squares (WLS) method is referred to as generalized least squares (GLS).

the correlation matrix, estimating a single factor with maximum likelihood extraction, and the printing of the reproduced and residual correlation matrices.

```
Chapter 6, Section §57, Data from Table 6.1
matrix data /variables=verbfreq bnaming oralword wordlist
/contents=corr mean stddev /format=full /n=200.
begin data.
```

1.00	0.51	0.52	0.43
0.51	1.00	0.40	0.54
0.52	0.40	1.00	0.50
0.43	0.54	0.50	1.00
11.1	11.9	23.6	4.3
5.4	3.0	11.8	2.0

```
end data.
factor /matrix=in(corr=*) /criteria=factors(1)
/extraction=ml /print=initial extraction repr.
```

The factor command produces a variety of output. We concentrate first on the communalities. The extraction communalities are the estimated proportion of the variance in each item that is explained by the common factor.

Communalities

	Initial	Extraction
verbfreq	0.383	0.486
bnaming	0.387	0.492
oralword	0.365	0.454
wordlist	0.391	0.502

Next, we have the factor loadings. With a single factor, these are the standardized regression coefficients, predicting each item by using the factor as the independent variable. In this case, they are also equal to the correlations between the factor and each item.

Factor Matrix[a]

	Factor 1
Verbfreq	0.697
Bnaming	0.701
Oralword	0.674
Wordlist	0.709

We now have the reproduced correlations based on the single-factor model, and the associated correlation residuals, which are the differences between the original correlations and

the model-implied ones. Note that the diagonal of the reproduced correlation matrix also contains the final communality estimates seen earlier.

Reproduced Correlations					
		verbfreq	**bnaming**	**oralword**	**wordlist**
Reproduced Correlation	verbfreq	**0.486**[a]	0.489	0.470	0.494
	bnaming	0.489	**0.492**[a]	0.473	0.497
	oralword	0.470	0.473	**0.454**[a]	0.478
	wordlist	0.494	0.497	0.478	**0.502**[a]
Residual[b]	verbfreq		0.021	0.050	−0.064
	bnaming	0.021		−0.073	0.043
	oralword	0.050	−0.073		0.022
	wordlist	−0.064	0.043	0.022	

[a] Reproduced communalities

[b] Residuals are computed as observed minus reproduced correlations. There are three (50.0%) nonredundant residuals with absolute values equal to or greater than 0.05.

CD2 *Factor analysis of the first tapping test using SPSS (Communalities, factor loadings, reproduced correlations, and factor scoring coefficients.)* Observed data for four variables from 12 individuals is given in Table 6.5 of Section §61. These data come from Petr's initial administration of his "tapping test" to twelve students in his class. Here, we use SPSS' data list command to read the data. The following factor command uses maximum likelihood factor extraction and requests the univariate descriptive statistics and the original correlations along with the factor loadings and communalities in the output. It also requests the factor scoring coefficients using Bartlett's method, and saves the associated Bartlett (1938) method factor score values on the active dataset.

```
Chapter 6, Section §61, Data from Table 6.5
data list list /person x1 to x4.
begin data.
```

1	12	15	28	30
2	22	23	48	37
3	12	15	20	29
4	15	21	35	42
5	17	21	27	26
6	19	23	37	41
7	17	24	39	37
8	21	25	40	45
9	12	18	33	26
10	17	17	29	33
11	16	21	29	33
12	12	15	28	30

```
end data.
factor variables=x1 x2 x3 x4
/extraction=ml
/print=initial extraction repr univariate correlation fscore
/save=bart(1).
```

We first have the means and standard deviations for the variables, followed by the estimated correlations among the items.

Descriptive Statistics

	Mean	Std. Deviation	Analysis *N*
*x*1	16.0000	3.54196	12
*x*2	19.8333	3.68864	12
*x*3	32.7500	7.48483	12
*x*4	34.0833	6.31676	12

Correlation Matrix

		*x*1	*x*2	*x*3	*x*4
Correlation	*x*1	1.000	0.842	0.765	0.666
	*x*2	0.842	1.000	0.772	0.703
	*x*3	0.765	0.772	1.000	0.664
	*x*4	0.666	0.703	0.664	1.000

The communalities and factor loadings are then produced.

Communalities

	Initial	Extraction
*x*1	0.745	0.819
*x*2	0.766	0.857
*x*3	0.659	0.711
*x*4	0.537	0.572

Factor Matrix[a]

	Factor 1
*x*1	0.905
*x*2	0.926
*x*3	0.843
*x*4	0.757

As before, the reproduced and residual correlations are then given.

Reproduced Correlations					
		*x*1	*x*2	*x*3	*x*4
Reproduced Correlation	*x*1	**0.819**[a]	0.838	0.763	0.685
	*x*2	0.838	**0.857**[a]	0.781	0.701
	*x*3	0.763	0.781	**0.711**[a]	0.638
	*x*4	0.685	0.701	0.638	**0.572**[a]
Residual[b]	*x*1		0.004	0.002	−0.018
	*x*2	0.004		−0.009	0.002
	*x*3	0.002	−0.009		0.026
	*x*4	−0.018	0.002	0.026	

[a] Reproduced communalities.
[b] Residuals are computed as observed minus reproduced correlations. There are 0 (0.0%) nonredundant residuals with absolute values greater than 0.05.

And finally, the Bartlett (1938) method factor scoring coefficients are produced. These coefficients are multiplied by the corresponding standardized variable values to give the Bartlett method factor score estimates for each individual.

Factor Score Coefficient Matrix	
	Factor 1
*x*1	0.348
*x*2	0.453
*x*3	0.204
*x*4	0.124

CD3 *Factor analysis of the second tapping test using SPSS (Z-scores and estimated factor scores.)* Individual data from a second administration of Petr's four-item tapping test to the same 12 students on a subsequent occasion are given in Table 6.7 of Section §61. The descriptives command with the /save option is used to calculate the corresponding *z*-scores for the four original variables. The formats and variable width commands are used to set the formatting of the variables for display later.

```
Chapter 6, Section §61, Data from Table 6.7
data list free / person x1 x2 x3 x4.
begin data.
```

1	10	17	30	29
2	23	23	50	37
3	12	14	19	35
4	15	24	35	42
5	18	21	29	26
6	22	27	37	40
7	18	24	38	36
8	22	25	45	45
9	11	18	32	28
10	16	16	29	32
11	18	22	28	33
12	14	14	28	33

```
end data.
descriptives x1 x2 x3 x4 /save.
formats person x1 x2 x3 x4 (f6.0) zx1 zx2 zx3 zx4 (f6.2).
variable width x1 to x4 zx1 to zx4 (6).
factor variables=x1 x2 x3 x4
/extraction=ml
/print=initial extraction repr univariate correlation fscore
/save=bart(1).
list.
```

As before, the factor command produces a variety of statistics from the factor analysis for the second occasion data, including the Bartlett (1938) method factor scoring coefficients for the second occasion data.

Factor Score Coefficient Matrix

	Factor 1
*x*1	0.387
*x*2	0.403
*x*3	0.275
*x*4	0.126

The list command produces the following listing of the original variables, and their *z*-score counterparts, along with the estimated factor score for each person.

```
person    x1    x2    x3    x4    Zx1    Zx2    Zx3    Zx4    FAC1_1

1         10    17    30    29   -1.50  -0.77  -0.40  -0.99  -1.12456
2         23    23    50    37    1.47   0.58   2.00   0.41   1.40108
3         12    14    19    35   -1.05  -1.44  -1.72   0.06  -1.44930
4         15    24    35    42   -0.36   0.80   0.20   1.28   0.40008
```

5	18	21	29	26	0.32	0.13	-0.52	-1.51	-0.15620
6	22	27	37	40	1.24	1.48	0.44	0.93	1.31090
7	18	24	38	36	0.32	0.80	0.56	0.23	0.63178
8	22	25	45	45	1.24	1.03	1.40	1.80	1.50456
9	11	18	32	28	-1.28	-0.54	-0.16	-1.16	-0.90206
10	16	16	29	32	-0.13	-0.99	-0.52	-0.47	-0.65176
11	18	22	28	33	0.32	0.35	-0.64	-0.29	0.05529
12	14	14	28	33	-0.59	-1.44	-0.64	-0.29	-1.01981

CD4 *Factor analysis using SPSS, comparing the first and second tapping test data and also calculating a correlation coefficient between the two sets of factor scores for the two tapping tests as an estimate of test reliability*. Finally, we combine the data for both testing occasions as found in Tables 6.5 and 6.7, where the four variables on the first occasion are designated as *x*11, *x*12, *x*13, and *x*14. The first digit (1) indicates the (first) occasion, and the second digit indicates the variable (1–4). A similar pattern is used for the second occasion, designating the variables *x*21, *x*22, *x*23, and *x*24. Correlations are estimated for all eight variables across both occasions, and corresponding *z*-scores are created for each of the eight variables. Separate factor models are created for each occasion each time saving the occasion-specific factor scores. Finally, a correlation coefficient is calculated between the pre- and post-factor scores as an estimate of reliability that is appended to the end of Table 6.7 under the title "Generic reliability of the latent score approximated by the model–remodel correlation: $r_{ff} = 0.986$." Now re-run th

```
Chapter 6, Section §61, Data from Table 6.5 and Table 6.7.
data list free / person
x11 x12 x13 x14 x21 x22 x23 x24.
begin data.
```

1	12	15	28	30	10	17	30	29
2	22	23	48	37	23	23	50	37
3	12	15	20	29	12	14	19	35
4	15	21	35	42	15	24	35	42
5	17	21	27	26	18	21	29	26
6	19	23	37	41	22	27	37	40
7	17	24	39	37	18	24	38	36
8	21	25	40	45	22	25	45	45
9	12	18	33	26	11	18	32	28
10	17	17	29	33	16	16	29	32
11	16	21	29	33	18	22	28	33
12	12	15	28	30	14	14	28	33

```
end data.

correlations x11–x24.
descriptives x11 x12 x13 x14 x21 x22 x23 x24 /save.

factor variables=x11 x12 x13 x14
/extraction=ml
```

```
/print=initial extraction repr univariate correlation fscore
/save=bart(1 pre).

factor variables=x21 x22 x23 x24
/extraction=ml
/print=initial extraction repr univariate correlation fscore
/save=bart(1 post).

correlations pre1 with post1.
```

The correlations shown in **bold** are the correlations between the same item at the first and second occasions.

Correlations

	*x*11	*x*12	*x*13	*x*14	*x*21	*x*22	*x*23	*x*24
Pearson Correlation								
*x*11	1	0.842	0.765	0.666	**0.955**	0.742	0.803	0.533
*x*12	0.842	1	0.772	0.703	0.851	**0.938**	0.771	0.574
*x*13	0.765	0.772	1	0.664	0.707	0.725	**0.979**	0.537
*x*14	0.666	0.703	0.664	1	0.659	0.750	0.678	**0.943**
*x*21	**0.955**	0.851	0.707	0.659	1	0.764	0.726	0.560
*x*22	0.742	**0.938**	0.725	0.750	0.764	1	0.715	0.597
*x*23	0.803	0.771	**0.979**	0.678	0.726	0.715	1	0.547
*x*24	0.533	0.574	0.537	**0.943**	0.560	0.597	0.547	1

The Bartlett (1938) method factor scoring coefficients for the first and second occasions are shown in the following two tables. The coefficients are quite similar but not identical.

Factor Score Coefficient Matrix

	Factor 1
*x*11	0.348
*x*12	0.453
*x*13	0.204
*x*14	0.124
*x*21	0.387
*x*22	0.403
*x*23	0.275
*x*24	0.126

The final correlation between the Bartlett (1938) method factor scores from the first and second occasions is 0.986.

Correlations	
Pearson Correlation	
	post1 BART factor score 1 for analysis
pre1 BART factor score 1 for analysis	0.986

CD5 *A computational caveat regarding the problem with factoring matrices with extremely low correlation coefficients*. Petr gives an instructive example (not included in any of the chapters of this book) of the results of factoring a correlation matrix with very small correlations. Here, the matrix data command in SPSS reads a lower triangular symmetric correlation matrix, and the factor command provides the usual statistics.

```
Chapter 6 addendum
matrix data /variables=x1 x2 x3 x4
/contents=corr /n=200.
begin data.
```

1.000			
0.091	1.000		
0.121	0.090	1.000	
0.093	0.109	0.119	1.000

```
end data.
factor /matrix=in(corr=*) /extraction=ml
/print=initial extraction correlation repr.
```

Note that the communalities and factor loadings are quite small, and even though the residuals are also very small, the utility of such a factor solution is problematic. Even though the residual correlation matrix seems to show partial uncorrelatedness, the analysis obviously could not be used to establish an LCF.

Communalities		
	Initial	**Extraction**
*x*1	0.026	0.098
*x*2	0.023	0.083
*x*3	0.031	0.125
*x*4	0.029	0.112

Factor Matrix[a]	
	Factor 1
*x*1	0.313
*x*2	0.289
*x*3	0.354
*x*4	0.335

		Reproduced Correlations			
		*x*1	*x*2	*x*3	*x*4
		Reproduced Correlations			
		*x*1	*x*2	*x*3	*x*4
Reproduced Correlation	*x*1	**0.098**[a]	0.090	0.111	0.105
	*x*2	0.090	**0.083**[a]	0.102	0.097
	*x*3	0.111	0.102	**0.125**[a]	0.118
	*x*4	0.105	0.097	0.118	**0.112**[a]
Residual[b]	*x*1		0.001	0.010	−0.012
	*x*2	0.001		−0.012	0.012
	*x*3	0.010	−0.012		0.001
	*x*4	−0.012	0.012	0.001	

[a] Reproduced communalities.
[b] Residuals are computed as observed minus reproduced correlations. There are 0 (0.0%) nonredundant residuals with absolute values greater than 0.05.

Computational demonstrations using Stata commands

CD6 *Factor analysis of the correlation matrix from the analysis of the four verbal tests of Table 6.1 using Stata (Uniquenesses, factor loadings, and reproduced correlations.)* For the data from Table 6.1 in Section §57, we can read the (full) correlation matrix using Stata's `matrix` command. We can then perform a factor analysis of the correlations using the `factormat` command, specifying a sample size of 200 and extracting a single factor using maximum likelihood estimation. The `estat residuals` command gives the reproduced and residual correlation matrices.

```
Chapter 6, Section 6.1.3, Data from Table 6.1
clear
matrix C = /// Input the correlation matrix
(1.00,  0.51,  0.52,  0.43 \ ///
0.51, 1.00,  0.40,  0.54 \ ///
0.52,  0.40, 1.00,  0.50 \ ///
0.43,  0.54,  0.50, 1.00)
factormat C, n(200) fac(1) ml /// One factor via maximum
likelihood
names(verbfreq bnaming oralword wordlist) // variable names
estat residuals, fitted       // Reproduced and residual matrices
```

Stata gives the same factor loadings as SPSS but gives the variable uniquenesses (one minus the communality) rather than the communalities themselves.

Factor loadings (pattern matrix) and unique variances

Variable	Factor1	Uniqueness
verbfreq	0.6971	0.5141
bnaming	0.7014	0.5080
oralword	0.6738	0.5460
wordlist	0.7088	0.4976

The reproduced or fitted correlations and the correlation residuals
Fitted ("reconstructed") values for correlations

Variable	verbf~q	bnaming	oralw~d	wordl~t
verbfreq	1.0000			
bnaming	0.4889	1.0000		
oralword	0.4696	0.4726	1.0000	
wordlist	0.4941	0.4972	0.4776	1.0000

Raw residuals of correlations (observed-fitted)

Variable	verbf~q	bnaming	oralw~d	wordl~t
verbfreq	−0.0000			
bnaming	0.0211	0.0000		
oralword	0.0504	−0.0726	0.0000	
wordlist	−0.0641	0.0428	0.0224	−0.0000

CD7 *Factor analysis of the first tapping test using Stata (Communalities, factor loadings, reproduced correlations, and factor scoring coefficients.)* Using the individual data from Table 6.5 in Section §61, the input command is used to read the data.

```
Chapter 6, Section 8.4.1, Data from Table 6.5.
clear
input person x1-x4 // Input the data
 1    12   15   28    30
 2    22   23   48    37
 3    12   15   20    29
 4    15   21   35    42
 5    17   21   27    26
 6    19   23   37    41
 7    17   24   39    37
 8    21   25   40    45
 9    12   18   33    26
10    17   17   29    33
11    16   21   29    33
12    12   15   28    30
end
```

```
factor x1-x4, fac(1) ml // One factor via Maximum Likelihood
estat residuals, fitted // Reproduced and residual matrices
predict bart, bartlett  // Factor scores and scoring coefficients
list, separator(0)      // List data and factor scores
```

The factor command estimates a single factor model for the four items using maximum likelihood estimation, the estat residuals command gives the reproduced and residual correlation matrices, the predict bart command gives the Bartlett (1938) method factor scoring coefficients and factor scores, and the list command provides a listing of the data and the factor scores.

Factor loadings (pattern matrix) and unique variances

Variable	Factor 1	Uniqueness
*x*1	0.9048	0.1813
*x*2	0.9259	0.1426
*x*3	0.8431	0.2892
*x*4	0.7566	0.4275

Reproduced and residual correlations
Fitted ("reconstructed") values for correlations

Variable	*x*1	*x*2	*x*3	*x*4
*x*1	1.0000			
*x*2	0.8378	1.0000		
*x*3	0.7628	0.7806	1.0000	
*x*4	0.6846	0.7006	0.6379	1.0000

Raw residuals of correlations (observed-fitted)

Variable	*x*1	*x*2	*x*3	*x*4
*x*1	−0.0000			
*x*2	0.0041	0.0000		
*x*3	0.0019	−0.0085	0.0000	
*x*4	−0.0182	0.0024	0.0259	−0.0000

Factor scoring coefficients
Scoring coefficients (method = Bartlett)

Variable	Factor 1
*x*1	0.34839
*x*2	0.45328
*x*3	0.20352
*x*4	0.12355

Factor scores

	Person	*x*1	*x*2	*x*3	*x*4	bart
1.	1	12	15	28	30	−1.196411
2.	2	22	23	48	37	1.451005
3.	3	12	15	20	29	−1.433494
4.	4	15	21	35	42	0.2610279
5.	5	17	21	27	26	−0.0727227
6.	6	19	23	37	41	0.9350638
7.	7	17	24	39	37	0.8373709
8.	8	21	25	40	45	1.537365
9.	9	12	18	33	26	−0.7700398
10.	10	17	17	29	33	−0.3729671
11.	11	16	21	29	33	0.0202123
12.	12	12	15	28	30	−1.196411

CD8 *Factor analysis using Stata, comparing the first and second tapping test data and also calculating a correlation coefficient between the two sets of factor scores for the two tapping tests as an estimate of test reliability*. Combining the data from Tables 6.5 and 6.7, the following syntax reads pretest and posttest data for 12 individuals, estimating separate factor models for the pretest and posttest along with their factor scoring coefficients and estimated factor scores. Finally, the pretest and posttest factor scores are correlated.

```
Chapter 6, Section 8.5.1, Data from Table 6.5 and Table 6.6.
clear
input person x11-x14 x21-x24 // Input the data
 1    12   15   28   30   10   17   30   29
 2    22   23   48   37   23   23   50   37
 3    12   15   20   29   12   14   19   35
 4    15   21   35   42   15   24   35   42
 5    17   21   27   26   18   21   29   26
 6    19   23   37   41   22   27   37   40
 7    17   24   39   37   18   24   38   36
 8    21   25   40   45   22   25   45   45
 9    12   18   33   26   11   18   32   28
10    17   17   29   33   16   16   29   32
11    16   21   29   33   18   22   28   33
12    12   15   28   30   14   14   28   33
end

correlate x11-x24          // Test inter-correlations
factor x11-x14, fac(1) ml // Pre one factor via Maximum Likelihood
```

```
predict pre, bartlett      // Pre factor scores and scoring
coefficients
factor x21-x24, fac(1) ml // Post one factor via Maximum
Likelihood
predict post, bartlett    // Post factor scores and scoring
coefficients
correlate pre post        // Pre-post factor score correlation
```

We first show the intercorrelations among all eight variables.

	x11	x12	x13	x14	x21	x22	x23	x24
x11	1.0000							
x12	0.8419	1.0000						
x13	0.7647	0.7722	1.0000					
x14	0.6664	0.7029	0.6638	1.0000				
x21	0.9554	0.8508	0.7066	0.6587	1.0000			
x22	0.7422	0.9383	0.7249	0.7503	0.7636	1.0000		
x23	0.8035	0.7705	0.9789	0.6778	0.7263	0.7145	1.0000	
x24	0.5334	0.5739	0.5366	0.9434	0.5596	0.5967	0.5471	1.0000

Next, we have the factor loadings and unique variances for the pretest variables x11–x14.
Factor loadings (pattern matrix) and unique variances

Variable	Factor 1	Uniqueness
x11	0.9048	0.1813
x12	0.9259	0.1426
x13	0.8431	0.2892
x14	0.7566	0.4275

We then have the Bartlett (1938) method factor scoring coefficients.
Scoring coefficients (method = Bartlett)

Variable	Factor 1
x11	0.34839
x12	0.45328
x13	0.20352
x14	0.12355

Next are the factor posttest factor loadings and uniquenesses, followed by the posttest factor scoring coefficients.

Factor loadings (pattern matrix) and unique variances

Variable	Factor 1	Uniqueness
*x*21	0.8714	0.2407
*x*22	0.8761	0.2325
*x*23	0.8245	0.3202
*x*24	0.6624	0.5612

Scoring coefficients (method = Bartlett)

Variable	Factor 1
*x*21	0.38681
*x*22	0.40249
*x*23	0.27507
*x*24	0.12611

```
correlate pre post // Pre-post factor score correlation
(obs = 12)
| pre post
-------------+------------------
pre | 1.0000
post | 0.9857 1.0000
```

References

Armstrong, J. S. (1967). Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine. *American Statistician*, 21(5), 17–21.

Armstrong, J. S. (1970). How to avoid exploratory research. *Journal of Advertising Research*, 10, 27–30.

Bartlett (1938). Methods of estimating mental factors. *Nature*, 141, 609–611.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.

Brown, B. L., Hendrix, S. B., Hedges, D. W., & Smith, T. B. (2012). *Multivariate analysis for the biobehavioral and social sciences: A graphical approach*. New York: John Wiley & Sons.

Burt, C. (1940). *The factors of the mind*. London University Press.

Costner, H. L. (1969). Theory, deduction, and rules of correspondence. *The American Journal of Sociology*, 75, 245–263.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Emmett, W. G. (1949). Factor analysis by Lawley's method of maximum likelihood. *British Journal of Psychology (Statistical Section)*, 2, 90–97.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. R. Lazarfeld (ed.), *Mathematical thinking in the social sciences*, Glencoe, IL: The Free Press.

Guttman, L. (1955). The determinacy of factor score matrices, with implications for five other basic problems of common-factor theory. *British Journal of Mathematical Statistical Psychology*, 8, 65–81.

Harman, H. H. (1960). *Modern factor analysis*. Chicago: The University of Chicago Press.

Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3), 209–229.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor analysis as a statistical method*, *2nd ed*., Butterworths, London.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA, Addison-Wesley.

Maraun, M. (1996). The claims of factor analysis. *Multivariate Behavioral Research*, 31, 673–690.

McDonald (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, 23, 1–21.

McDonald, R. P. (1975). Descriptive axioms for common factor theory, image theory, and component theory. *Psychometrika*, 40, 137–152.

McDonald, R. P. (1977). The indeterminacy of components and the definition of common factors. *British Journal of Mathematical and Statistical Psychology*, 30,165–176.

McDonald, R. P. (1978a). Generalizability in factorable domains: "Domain validity and generalizability". *Educational and Psychological Measurement*, 38, 75–79.

McDonald, R. P. (1982). Some alternative approaches to the improvement of measurement in education and psychology: Fitting latent trait models. In D. Spearitt (ed.), *The improvement of measurement in education and psychology*. ACER, 213–237.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, 6, 97–103.

Mulaik, S. (2005). Looking back on the indeterminacy controversies in factor analysis. In A. Maydeu-Olivares, & J. J. McArdle (eds.), *Contemporary psychometrics, a festschrift for Roderick P. McDonald*. Mahwah, NJ: Lawrence Erlbaum. 173–206.

Mulaik, S., & McDonald, R. P. (1978). The effect of additional variables on factor indeterminacy in models with a single common factor. *Psychometrika*, 43, 177–192.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6* 2 (11) 559–572.

Rencher, A. C. (2002). *Methods of multivariate analysis*. New York: John Wiley & Sons.

Thomson, G. H. (1951). *The factorial analysis of human ability*, *5th ed*., London University Press.

Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, 41(1), 1–32. https://doi.org/10.1037/h0075959

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

7

Analysis of Multiple Factors

CHAPTER MENU

7.1 Modeling Multiple and Hierarchical Constructs, 277
7.1.1 (§63) Multiconstruct Theories Including Two-Layer Structures, 278
7.1.1.1 Bi-factor Models, 278
7.1.1.2 Hierarchical Models, 278
7.1.1.3 Dimensionality and Higher-Order Constructs, 280
7.1.1.4 Constructs with Multiple Layers, 280
7.1.2 (§64) Homogeneity and Dimensionality in One-Layer Theories of Abilities: *Independent Cluster Structure (ICS)* Compared to *Simple Structure*, 282
7.1.2.1 Independent Cluster Structure, 282
7.1.2.2 Factor Rotation and Simple Structure, 284
7.1.3 (§65) Unidimensionality vs. Multidimensionality on One Layer, an Example, 285
7.2 Confirmatory and Exploratory, 287
7.2.1 (§66) The Confirmatory Approach, 287
7.2.2 (§67) Confirmatory vs. Exploratory Simple Structure, 288
7.2.2.1 Simple Structure, 290
7.2.2.2 Exploratory Common Factor Analysis, 291
7.2.2.3 Factor Rotation, 292
7.2.3 (§68) Stronger vs. Weaker Support for Content Homogeneity: By Unidimensionality or ICS, 294
7.2.3.1 Higher-Order Factor Analysis, 295
7.2.3.2 Equivocal Indicators and Factorial Complexity, 296
7.2.4 (§69) "Hidden Equivocality" in ICS Structures and the So-Called Frame Problem, 296
7.2.4.1 Indicator Multidimensionality, 298
7.2.4.2 The "Frame Problem", 299
7.2.5 (§70) Dimensionality by the LC Factor *Model* vs. PCA "Factor *Analysis*"—Example, 300
7.2.5.1 Explaining Correlations vs. Explaining Variance-1, 300
7.2.5.2 Meaninglessness of "Communalities" in PCA-2, 303
7.2.5.3 What "Variance" Is Explained by PCA-3, 303
7.2.5.4 Deterministic Component Scores vs. Factor Score Indeterminacy, 305
7.3 Summary, 306
7.4 Study Questions, 307
References, 313

7.1 Modeling Multiple and Hierarchical Constructs

Domains of interest in human sciences research and instrument development seldom have just one factor. As in the fluent oral reading example earlier, the oral reader must know the word in order to smoothly read a phrase containing it. Such dependencies of correlated and necessary factors are

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

common, and if domain objectives include knowing what the names of terms are, that pesky linguistic factor is always hanging around. Many other domain analyses will quickly find that complex behaviors are composed of multiple parts and just cannot be considered to be one-dimensional. It is likely that not all interesting structures have been studied and named, but a number of them have, and several are shown graphically and discussed in Sections §63–§65.

7.1.1 (§63) Multiconstruct Theories Including Two-Layer Structures

In earlier sections, we have touched briefly on some examples of structural theories. The single-factor model of verbal ability of four verbal subtests of the CERAD was a very simple case while a historical example of the development of structural theories of intelligence presented somewhat more complicated theories. For didactic illustration of latent variable (LV) modeling, the numerical examples in Chapter 6 dealt with unidimensional cases such as general verbal ability or Spearman's general *g* theory of intelligence. Such simple *single-construct theories* with a general *construct*, which is common to all empirical indicators may be represented graphically in terms of their LCF models by a *path diagram* with one latent *general factor*. An early theory competing with the general Spearman's *g* was Thurstone's (1935, 1938) theory of primary mental abilities which expanded the focus to multiple constructs. It was a "one-layer theory" of a *combinatory* character, expanded to include six abilities: *M*—memory, *R*—reasoning, *V*—verbal comprehension, *W*—word fluency, *S*—spatial ability, and *N*—numerical ability.

7.1.1.1 Bi-factor Models

Holzinger's bi-factor theory (e.g. Holzinger & Swineford, 1937) would not be considered a genuinely two-level hierarchical theory, yet that depends on how one defines or understands the notion of "hierarchy." This bi-factor theory kept Spearman's *g* as general ability, common to the whole set of indicators, while each of the six ability groups, *M*, *R*, *V*, *W*, *S*, and *N*, was "common" to a small subgroup of indicators. This is the logical source where the notion of a "common factor" in the LCF model comes from. From this point of view, the latent common factors (LCFs), which are modeling the abilities in the bi-factor theory, are all in the same order/layer of hierarchy. They belong to the lowest theoretical layer, closest to the empirical indicators to which each of the LC factors is common. They are common to at least any *two* indicators, that is, common in a meaningful minimum sense. Hence, though belonging to the same order of hierarchy, they each have their own *referent generality* in the meaning of Coan (1964). Perhaps, we could call this type of structural theory "quasi-hierarchical," as shown in Figure 7.1a,b.

7.1.1.2 Hierarchical Models

A *genuine hierarchical* organization should ideally correspond to a situation where concepts of a higher-order layer are primarily related to concepts of a lower-order. Hence, typically, *a higher-order "construct" is common to several lower-order constructs*. It does not mean that higher layer constructs must not be directly related to the lowest layer of empirical indicators at all. Nevertheless, if a higher-order construct were directly related to *all* indicators, it would become one of the common lower-order constructs, namely a general one, an analog of *g*. Then, we would have a contradiction in the logical distinction of what is higher and/or lower-order of the theoretical concepts. Bi-factor theory (Figure 7.1a,b), with *g* as a common ability of *larger generality*, contrasts with a

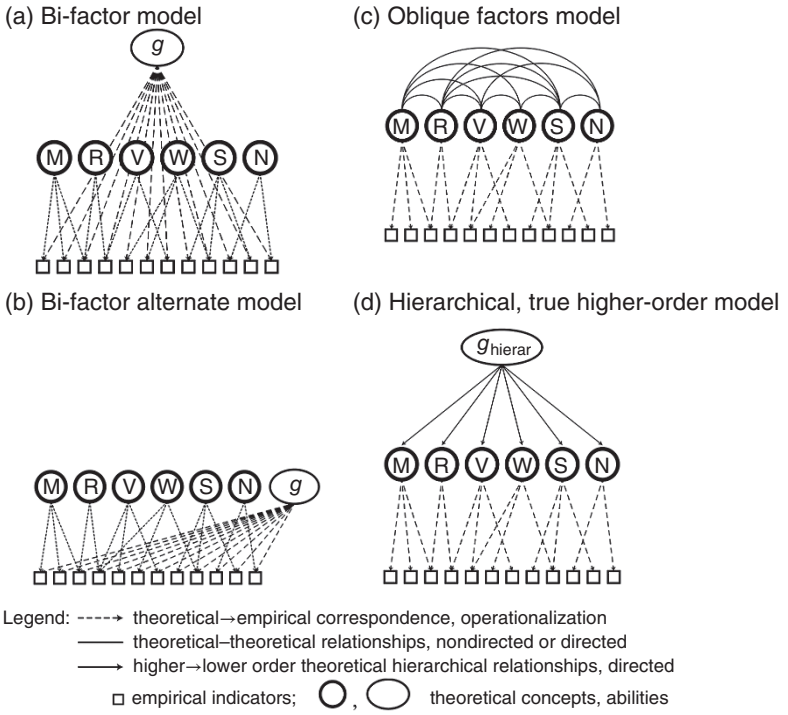


Figure 7.1 Hierarchy and generality: an illustration by historical theories of intelligence.

genuine hierarchical theory as shown in Figure 7.1d, with an illustrative $\mathbf{g}_{\text{hierar}}$ as a *concept of second order on a higher layer*.

Figure 7.1 presents four situations. A comparison of pictures (a) and (b) indicates that the bi-factor theory has a "quasi-hierarchical" structure only: the seemingly higher layer of *g* in (a) is due only to placing the general ability graphically above the group abilities. In picture (b), it shows symbolically that *g*, as well as *M*, *R*, *V*, *W*, *S*, and *N*,[1] could all be *common* abilities of the same order of hierarchy, though *g* has a higher generality, actually the highest possible in this case since it includes all of the empirical indicators. Its contrasting counterpart of a genuine hierarchy is in picture (d) with *M*, *R*, *V*, *W*, *S*, and *N* in the role of first-order constructs and g_{hierar} as a second-order one. Picture (c) represents the *one-layer theory* of six constructs that are mutually interrelated. In picture (d), g_{hierar} is related to the six theoretical concepts in an analogous way as *g* is related to the empirical indicators. Hence, we may say that g_{hierar} is a hierarchical counterpart of Spearman's General Intelligence as a *super-construct* but with generality on a higher layer of hierarchy. Of course, it is conceptually different from bi-factor theory in that it is separated from the empirical indicators by two steps rather than one. However, it seems that a certain version of a general characterization of intellectual abilities is still not fully abandoned, actually, it is supported by the usual

1 As stated in the second paragraph of this chapter "MRVWSN" stands for the initial letters of "Memory," "Reasoning," "Verbal comprehension," "Word fluency," "Spatial ability," and "Numerical ability" from Thurstone's (1938) theory of mental abilities that historically competed with Spearman's (1904) *g* theory of general intelligence.

practice of *IQ* being used as a scalar, the one-valued "quotient" on a unidimensional scale. On the other hand perhaps in this, and in many other assessments, clinical diagnostic methods are influenced too much by an unscientific tendency to express complicated things in terms of naively oversimplistic numerical values, diagrams, and symbols.

From the point of view of *consistency of correspondence to the common "construct"* of several indicators, we find it certainly stronger in picture (a) of Figure 7.1 where the indicators are linked directly to *g*, the common construct. On the other hand, in the hierarchical theory (d), the consistency is considered weaker, and in a more complicated form, since the connection to g_{hierar} is mediated through the lower-order theoretical subconcepts, the six abilities. Moreover, it is necessary to consider that the lower-order concepts may have *varying weights* of importance in their combination connecting them to the higher layer. For example, memory and reasoning could have quite different weights in their relationship to g_{hierar}.

7.1.1.3 Dimensionality and Higher-Order Constructs

In addition, we meet the topic of dimensionality specifically related to a certain layer of hierarchy. If dimensionality were studied by the LCF model, that is, by the *hierarchical LCF model*, known also as the *higher-order LCF model*, we would have a more complex structural hypothesis. Although the battery of indicators is not unidimensional as a whole, the set of six first-order LC factors in Figure 7.1d is still unidimensional on the second-order layer. This situation is mentioned generally, namely, that it must be made clear *on which layer/order of hierarchy we hypothesize the unidimensionality*. If the hierarchical structure were tested by the old-fashioned two-stage LCF modeling approach then, in the first step, the six LC factors would be fitted to the correlation matrix of empirical indicators (tests) with the hypothesis that the factors are mutually correlated. If this first step were considered separately, it would correspond to testing the six-dimensional LCF structure of Thurstone's one-layer theory pictured in (c). In the second step, the 6×6 correlation matrix of the first-order factors would be separately tested for second-order layer unidimensionality, with a single LC factor representing the super-construct, g_{hierar}, which should "explain" the 6×6 correlations of the lower layer constructs sufficiently well.

In some areas, where a theory is not yet elaborated into a more sophisticated structural system, the two-layer structure of a hierarchical organization, similar to picture (d) in Figure 7.1, may be found quite often. There may exist a vague and very general conceptual area which is named by an umbrella term (like "Wellness"), and this unclear notion is hypothesized to be subdivided into several "components." However, from the point of view of general methodology of science, these "components" should be better presented as theoretical concepts of a lower-order layer, *subconstructs* of a *super-construct*. The latter then actually stands for an abbreviated but somewhat vague name of the whole conceptual domain.

7.1.1.4 Constructs with Multiple Layers

This situation almost prompts that the "components" of the general notion might be rather speculative and wishful mental "constructions." It seems that such a type of *two-layer* structure with one general notion with its "components" is typical for the early stage of a theory in many behavioral and social sciences. In this premature stage of development, hierarchical structure may be conceptualized as a rather unclear combination of two types: first, the bi-factor theory with the common general construct; and second, a hierarchical theory with the higher-order g_{hierar}. Sometimes the graphical representation does not even follow the basic rules for drawing a correct path diagram.

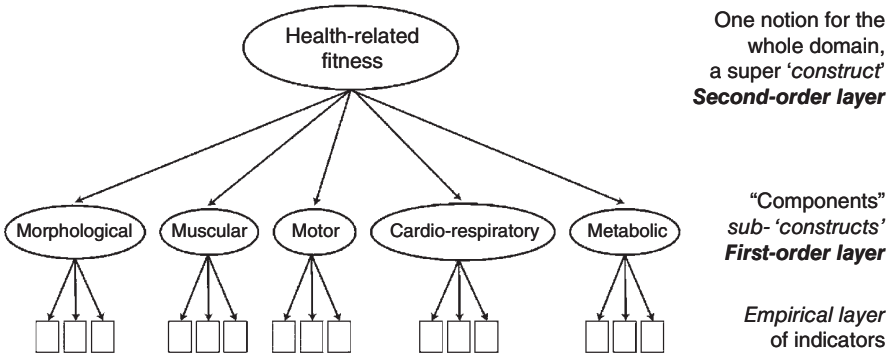


Figure 7.2 A schematic picture: an example of a *two-layer theory* with a general theoretical concept and its "components." (According to the theory by Bouchard, Shephard, & Stephens 1994, p.81 / Human Kinetics Publisher.)

Then an intuitive graphical outline of the theory may be presented without the arrows representing correspondence between theoretical concepts and empirical indicators since the indicators are not yet known. Actually, a precise knowledge of individual concrete indicators need not be of the main interest in such an early stage of more or less flexible and creative theorizing. We can see an example draft of a typical two-layer preliminary theory in the attempts to conceptualize the general notion of "*Health-Related Fitness*" with its lower layer of so-called "Fitness Components." Let us introduce Figure 7.2, the version from the recognized consensus book by Bouchard, Shephard, and Stephens (1994, p. 78).

There are many structural theories of this rather simple type. For example, Figure 7.3a presents one of the tentative theories about the conceptual structure of motor symptoms of patients with Parkinson's disease, as measured by 27 psychiatric clinical items in the battery known as the Motor Section of the UPDRS (Unified Parkinson's Disease Rating Scale). Figure 7.3a depicts the theory's four subordinate theoretical concepts under the umbrella of general motor symptoms of PD patients (Štochl et al. (2006), is a source publication). Figure 7.3b, a cognitive processes example, presents the theory and structure of attitudes toward physical activity and education, originally in Wear's (1951) inventory, the Attitudes Toward Physical Activity, many times modified by him as well as by other authors.

A question may be posed whether in this kind of theory, each of the first-order concepts (such as the abilities *M*, *R*, *V*, *W*, *S*, *N*) should be also considered unidimensional. This is a topic dealing with aspects of the so-called *simple structure*, originally a specific notion of the historical era of exploratory "factor analysis" (introduced by Thurstone). Nevertheless, the notion may be quite reasonably

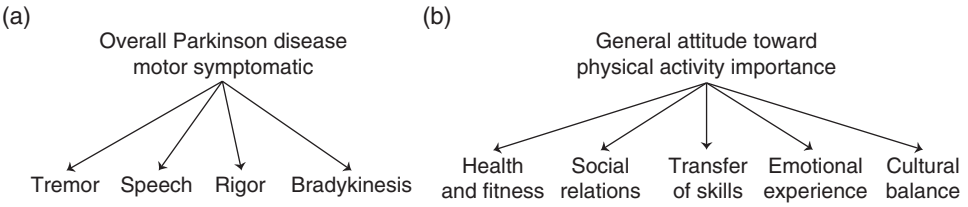


Figure 7.3 Examples of additional two-layer theories for understanding human functioning, with a general notion in "umbrella" position above its components (a) Is the factor structure of the UPDRS, (b) Is the factor structure of the ATPA.

borrowed and generally applied even to the structure of unformalized scientific theories and to the possible ways of organizing their theoretical concepts.[2]

7.1.2 (§64) Homogeneity and Dimensionality in One-Layer Theories of Abilities: *Independent Cluster Structure (ICS)* Compared to *Simple Structure*

In this section, we will deal only with one-layer theories and in particular with the comparison between *Simple Structure* one-layer theories and *Independent Cluster Structure (ICS)* one-layer theories. General approximate Simple Structure is exemplified by the lower-order layer of constructs shown in Figure 7.4b on the right, the diagram of Thurstone's six abilities theory. In contrast to this plan, Figure 7.4a on the left shows a diagram of a comparable one-layer structural hypothesis but this time modeling the modified 12-subtest CERAD battery as an example of ICS.

Obviously, there are major differences between these two ways of organizing the indicators and connecting them with the factors. While there are "crossovers" among the assignment of indicators to factors in simple structure (with some of the empirical indicators being assigned to two or more of the factors), in the ICS design, each of the three included subtests of the CERAD is measured by indicators which are clustered into groups which do not overlap, and no indicator is omitted. In the mathematical theory of sets, this is called *partitioning a set into separate subsets.*

7.1.2.1 Independent Cluster Structure

LV modeling uses a special technical terminology where the partitioned subsets of indicators are synonymously called *Independent Clusters*, and the whole complex an *ICS*. Note that this very special meaning of "independence" in LV-modeling jargon does not mean that indicators from different clusters should be mutually uncorrelated or that the LC factors representing separate clusters are uncorrelated. From the point of view of the individual indicators, it means the theory assumes that: *each indicator measures one and exclusively one theoretical concept*.[3]

Using other hybrid synonyms coming mostly from the history of LCF modeling, it also can be said that each indicator incorporated in ICS is a *univocal* indicator of just one—its own—theoretical concept. In other words, it has "*complexity*" equal to 1. The last jargon term should distinguish ICS indicators from the "conceptually complex" indicators that simultaneously indicate more than one construct or factor. For example, complexity equal to 2 means that such an indicator measures two constructs, etc.[4]

Of course, the ICS variant of a theoretical structure is conceived on the basis of the semantic content of the indicators with respect to their correspondence with theoretical concepts. It means that the *ICS clusters* are conceptualized (ideally) as *content homogeneous*. If this structure, ICS, is modeled by the LCF model, then each of the m independent clusters is represented by one latent factor, which is common to this cluster and only to it. It is usual that these LC factors are assumed to be possibly mutually correlated and that their inter-relationships might express that the concepts possibly belong to a common, more general, domain. Thus, we see that the independent clusters are

2 In many ways, latent common factors modeling is a quantitative/symbolic approach to understanding human functioning, including human cognition, through the creation and testing of mathematical models. Figure 7.2 is a simple example of the application of this LCF approach applied to human physical functioning. Figure 7.3 compares a model for physical functioning on the left side (a) to a model applied to the functioning of the mind, cognition, and attitudes (b) on the right.

3 That is, each empirical indicator can be said to have one and only one nonzero correlation (one "nonzero validity") defining the relationship between it and its factor, its construct.

4 Notice, however, that the "complexity" jargon is misleading since actually, a univocal indicator is not "complex" at all. Hence, its complexity should, by common sense, be called "zero complexity."

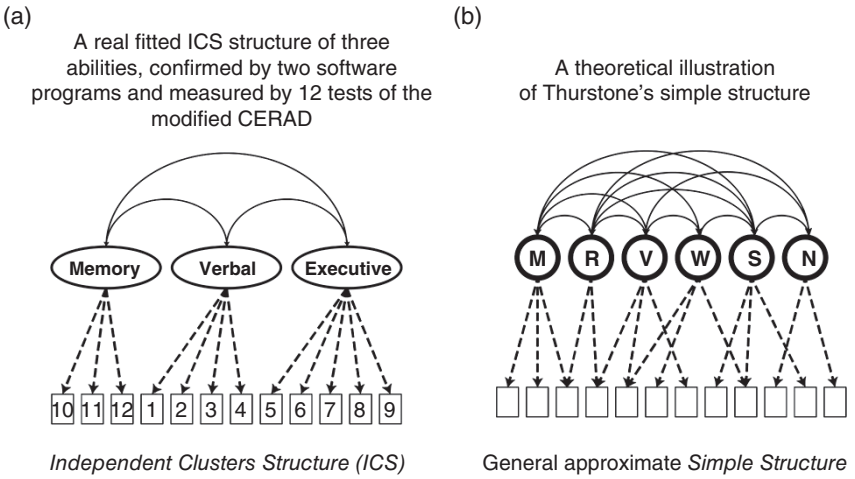


Figure 7.4 Two types of *simple structure*: (a) *ICS* vs. (b) *general approximate* non-ICS. (Subtests of the modified CERAD battery are re-ordered to exhibit ICS more transparently.)

usually quite reasonably assumed to be mutually "dependent" in the sense of correlation. These inter-relationships are represented by the nondirected arches above the first-order theoretical concepts in the examples in Figure 7.4.

The ideally disjunctive structure of indicators—ICS—may be too simplistic in some cases. Nevertheless, most recent theoretical conceptions, on which practical construction of tests and questionnaires are based, more or less automatically assume that each subtest or item should be conceptualized as a univocal indicator of just one theoretical concept. Such an exclusive assumption can be inadequate in situations when an indicator should reflect a mutual interplay of two (or more) theoretical concepts. For example, the performance, or another kind of outcome of a behavioral act, may be a result of mutual compensation, or even "competition" between two abilities. Certainly, such a type of *compensatory model* of the structure of a scientific theory should be conceived and appropriately justified on the ground of the corresponding substantive theory about the domain of behavior. Then it is important to take into consideration not only univocal indicators but also some indicators which reflect this mutual interplay, the indicators with higher complexity (i.e., with factor validity to more than just one LC factor). Obviously, in such a case, the ICS cannot be an appropriate paradigm for that particular substantive theory.

In the historical development of mathematical methods for so-called *exploratory* "*factor analysis*" (EFA), the predecessors of the LCF *model*, there were no options to pose any of the above-mentioned restrictions as an *a priori*-prescribed structural hypothesis representing the modeled substantive theory. For example, the idea of a theory based on ICS organization of constructs must enable a researcher to limit some LC factor validity coefficients—loadings—to be (exactly) zero in advance. Thurstone, in an effort to bring evidence for his one-layer theory of primary mental abilities, elaborated some roughly approximate numerical methods for the computation of loadings on more than a single (general) factor. One of them was the centroid method (section §57b of Chapter 6) that when combined with the Fundamental Factor Theorem (in section §57d of Chapter 6) enabled him to repeat its applications on a sequence of residual correlations to obtain loadings on additional types of factors. These methods were usually applied in two steps. In the first step, they produced a so-called initial or "unrotated basic solution." There was then a second step in which the basic solution was rotated for simplification.

7.1.2.2 Factor Rotation and Simple Structure

These "unrotated" loadings—factor validity coefficients—were far from optimum in guiding a researcher to distinguish and interpret which subtest/item might be an indicator for which factor. Therefore, some approximate and subjective methods were invented for the transformation of the "basic solution." This was carried out subjectively and step by step repeatedly on graph paper by "rotation" of pairs of (x, y) axes representing the factors. The target of these rotations was to obtain projections (coordinates) of the loadings of each indicator to be large on one factor and small on other factors, which was checked by inspection. This target was a little better specified by Thurstone's rules for approximating the so-called "Simple Structure." However, the rules remained somewhat ambiguous.[5]

In the present state of affairs, it is not necessary to offer a detailed list of his partial rules—their overall intuitive principle was more or less acceptably formulated by the first *mathematically optimal methods* for transformation into a simple structure of uncorrelated factors, namely the Quartimax method (Neuhaus & Wrigley, 1954) and the Varimax method (Kaiser, 1958), later programmed for computers. The term "Varimax" is more expressive in that it prompts the guiding principle: Maximize the variances of the loadings of an indicator over those of the factors. It offers better discrimination between large and small loadings. However, since there is no unique objective criterion about how to express the rather unclear intuitive definition, it is better to underscore that by applying these rotational methods to the basic exploratory solution, we may reach only an *approximate* Simple Structure. This approximation should be understood in the sense of bringing a given structure somehow closer to the unclear ideal of "best discrimination" of which indicators measure which factors. Depending on the intuition of mathematicians as authors, this idea can be formulated also by other additional criteria than just "Varimax." Thus, one may speculate which of the various rotational methods offers a "simpler simple structure" depending rather on easiness of the ex-post semantic interpretation of the factors. An example of this is shown numerically in Table 7.5 and diagrammatically in Figure 7.4a,b. On the other hand, there exist simple structures that are extensions of ICS and yet not so liberal as the Thurstonian approximate Simple Structure in its weakest meaning.

Imagine a case when a confirmatory strictly prescribed ICS were additionally augmented by a few additional prescribed loadings. Thus, the former ICS becomes an instance of a more general Simple Structure, but it can be unambiguously confirmed or rejected. No loadings are left to be "found by blind exploration" through the rotation. Thus, not every confirmatory prescribed Simple Structure needs to be only of the ICS type. There may exist several types of such simple structures between the two extremes. Perhaps, we could call such *verifiable prescribed structures* lying between the strict ICS and the general approximate Simple Structure "*nonapproximate simple structures*." In any case, the general term "simple structure" should *not* be confused with ICS, which is its special, narrower subcategory. Unfortunately, this kind of misunderstanding can be not too rarely found.

Problems with Thurstone's (1935, 1947) historical definition of exploratory *Simple Structure* were discussed in detail in McDonald (1985). As was mentioned above, we will just stay here with the somewhat unclear intuitive idea of "best discrimination" between indicators in a sense of which indicator more or less indicates which factor. Hence, Thurstone's conception allows a

5 Thurstone's centroid and related methods produced artificial variables, called "factors," which were not latent in the up-to-date proper meaning. Therefore, we omit the term "latent," and do not use abbreviations like LCF. The factors formerly were named after the corresponding roughly approximated computational method, such as "centroid factors," "principal factors," etc.

distribution of magnitudes of loadings of an indicator over several factors, which typically results in so-called

- one "*primary*" factor validity coefficient, the indicator's largest loading on a factor,
- "*secondary*" one, i.e. the next relatively high loading on another factor, and
- "*negligible*" loadings, usually the rest of those smaller than 0.30 in absolute value.

Probably, a reason for choosing 0.30 was based on the Fundamental Theorem (section §55): Contribution of the latter loading to its whole communality h_j^2 of the jth indicator was $a_j^2 = 0.3^2 = 0.09$, and the amount smaller than 0.10 seemed to be negligible, in those times. Even in some present articles, whose authors are still using certain obsolete methods, one can find results in tables where factor loadings smaller than $|0.30|$ were omitted. Though we need not suspect this publication strategy as an attempt to hide a bad fit or "not too simple" structure, such a Table with empty cells might be masking the fact that the method used was not the genuine confirmatory LCF model.

7.1.3 (§65) Unidimensionality vs. Multidimensionality on One Layer, an Example

As we saw in Table 6.1 of section §55 of Chapter 6, the cluster of verbal subtests in the CERAD battery was unidimensional.[6] Nevertheless, we know that the whole battery was constructed to cover several theoretical concepts. It was intended to measure three broad constructs, that is abilities, that are called memory, verbal, and executive. This theory did not assume hierarchical organization. Although we know that the original one-layer theory was three-dimensional let us—for a while—exaggerate a little the importance of parsimony, and check whether, perhaps, the whole modified battery of 12 subtests still might not be unidimensional. Perhaps, some additional competing theories would also favor the opinion that AD effects have mostly an overall common character. For a simplified illustration, let's have a look at only a part of the battery, taking just two clusters of subtests that are supposed to be indicators of two abilities, namely verbal and executive. Their 8×8 correlation matrix, showing all combinations of *internal manifest validity* between pairs of the subtests, is presented in Table 7.1.

In very simple cases, as in Table 7.1, it is sometimes possible to recognize certain structures of mutual correlations just by inspection of the correlation matrix. Usually, certain reorganization of rows and columns may be needed to form blocks of correlations and check whether the internal manifest validity between subtests is substantially higher inside the corresponding blocks according to the hypothesized structure. Then, such a correlation *matrix divided into blocks* called *submatrices* may reveal a *manifest structure* of correlations. However, in the example in Table 7.1, we can recognize on first sight that the manifest structure of the mutual internal validities of the subtests does not follow the hypothetically postulated blocks very well: Test 5 C-Praxis obviously belongs more to the first cluster than to the second where it is typically found, as revealed by the vector of its correlations with the other seven subtests. In a purely exploratory approach, such an inspection might serve for a specific application of the *inductive approach* to theorizing about *inferred* entities.

Nevertheless, in general, with the exception of quite underdeveloped branches of science, the hypothetico-deductive approach with its *postulated*, *logical* formation of theoretical concepts should be preferred. Unfortunately, we may occasionally witness a completely *blind* use of

6 This was also shown and explained in Table 3.2 of section §34 of chapter 3. In both places, the cluster of verbal subtests was used to demonstrate the rationale of the mathematics of calculating a latent factor variable *f*.

Table 7.1 Correlation matrix of eight subtests selected from the modified CERAD battery (Original negative correlations of subtests 6, 7 with the others were changed into positive—see text).

									Expected:	
j	Subtest x_j	1	2	3	4	5	6	7	8	Blocks according to clusters	
1.	Verbal fluency	1	**0.50**	**0.52**	**0.43**	*0.38*	0.19	0.20	0.22	**Higher** Manifest Validity	Lower
2.	Boston naming	**0.50**	1	**0.40**	**0.54**	*0.48*	0.10	0.10	0.20		
3.	Contr. oral word	**0.52**	**0.40**	1	**0.50**	*0.33*	0.18	0.20	0.28		
4.	Word list	**0.43**	**0.54**	**0.50**	1	*0.41*	0.27	0.24	0.38		
5.	Constr. praxis	*0.38*	*0.48*	*0.33*	*0.41*	*1*	*0.10*	*0.10*	*0.23*	Lower	**Higher** Manifest Validity
6.	Trails making A	0.19	0.10	0.18	0.27	*0.10*	1	**0.89**	**0.43**		
7.	Trails making B	0.20	0.10	0.20	0.24	*0.10*	**0.89**	1	**0.38**		
8.	Symbol digit	0.22	0.20	0.28	0.38	*0.23*	**0.43**	**0.38**	1		

exploratory *cluster analysis* without a glimpse of any hypothesis, and with no conceptual or substantive idea. An especially regrettable version of this kind happens if only the separate statistically significant correlations are taken into account and their designated cluster is used for speculations like "What might be a reason that these things happened to be correlated? What kind of sufficiently flexible name could be given to this gathering of variables?"

For illustration, in Table 7.1, we can see two blocks of higher correlations contrasting the lower correlations between clusters to support the hypothesis of two abilities. Furthermore, there seems to be a certain overlap, especially due to subtest 5, Constructional Praxis. Hence, the latter finding suggests that there might be a different latent structure than the originally assumed clear dimensionality of two. As a first thing, let's try to check and eliminate the possibility of unidimensionality. Actually, the procedure will be an analog of fitting the single verbal *g*-factor, subtests 1–4 of Table 7.2 that we successfully tested in Figure 6.1 of Chapter 6 for unidimensionality, but now extended with four subtests added from the executive *g*-factor, subtests 5–8 as shown in Table 7.2.

The evidence for unidimensionality in the LCF for the single verbal *g*-factor (subtests 1 to 4) in Table 6.1 was strong, with a comparative fit index (CFI) of 0.957, an acceptably low standardized root mean square residual (SRMR) of 0.038, and a maximum absolute value of residual correlation of $|r_{\mathrm{RES}}|$ max $= 0.071$. By contrast, we see in Table 7.2 that when we extend the correlation matrix to include subtests 5 through 8, the CFI drops to 0.50, the SRMR increases to 0.15, and the maximum residual correlation increases to $\lceil r_{res} \rceil \; max \; = 0.74$, with ten of the 28 residual correlations above 0.20. Confirmatory factor analysis (CFA) using Stata's sem procedure was used both for the analysis of unidimensionality for the four verbal subtests of the CERAD in Table 6.1 and also for the analysis of unidimensionality for the eight subtests, consisting of the verbal scale plus the executive scale of the CERAD. The comparison of the two analyses clearly suggests that the idea of unidimensionality of all eight subtests is untenable.

The findings in Table 7.2 are supporting what we assumed. Hence, the next step can be directed to the two-dimensional hypothesis of correlated ICS clusters—however, some of the higher residual correlations of subtest 5 in Table 7.1 indicate that this structure may not be too plausible either. Thus, our hypothesis about unidimensional structure, hopefully supporting the content homogeneity of macro-behavior of AD patients, should be rejected.

Table 7.2 The bad fit of LCF unidimensionality for the correlation matrix of Table 7.1 (estimated with Stata sem, maximum likelihood method).

Single factor model										
Estimated factor loadings		**Residual "unexplained" correlations**								
***j* Subtest x_j**	a_j	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Fit indices**
1. Verbal fluency	0.66	—			Fill in by symmetry					
2. Boston naming	0.66	0.07	—							
3. Contr. oral word	0.65	0.09	−0.03	—			r_{res}			$CFI = 0.50$
4. Word list	0.74	−0.06	0.05	0.02	—					$SRMR = 0.15$
5. Constr. praxis	0.55	0.02	*0.11*	−0.03	0.00	—				$\lceil r_{res} \rceil\ max = 0.74$
6. Trails making A	0.39	−0.07	*−0.16*	−0.08	−0.02	*−0.12*	—			
7. Trails making B	0.38	−0.05	*−0.15*	−0.05	−0.04	*−0.11*	***0.74***	—		*Residuals over 0.10*
8. Symbol digit	0.46	−0.08	*−0.11*	−0.02	0.04	−0.03	*0.25*	*0.20*	—	*are italicized*

Note: CFI = comparative fit index; SRMR = standardized root mean square residual.

7.2 Confirmatory and Exploratory

7.2.1 (§66) The Confirmatory Approach

The example analysis just presented in Table 7.2 is a practical instance of a general way to study the relational structure of a domain of research interest, namely the *confirmatory* approach. Contrary to the inductive approach with its exploratory "discovered" structures, this is one of the possible procedures of the *hypothetico-deductive* approach to concept formation and theory development outlined in Section §65. It presumes certain structures on mostly theoretical grounds and builds certain provisionally accepted theory for searching and finding structures that might appear or "emerge" from the data collected. In the confirmatory approach, some theoretically deducible and hypothetically presumed structures can be formulated as so-called *prescribed structural hypotheses*. These are to be tested by fitting them to data, generally to a covariance or correlation matrix, and if the fit is acceptable, such hypotheses about the structure of "constructs" may be provisionally accepted.

There are two typical forms for how a prescribed structural hypothesis can be formulated:

1) *Path diagrams*, briefly introduced in Figure 3.5 of section §35c, a graphical depiction with one-way arrows oriented to represent *regressions* of dependent on independent variables, →, for instance, of a manifest indicator on the latent factor by $f \rightarrow x$, or bi-directional interconnections that represent *correlations* between variables, ↔, for example, a correlation between two latent factors, say $f_1 \leftrightarrow f_2$. These are certain minimum rules for drawing path diagrams, perhaps the most important of which is the difference between bi-directional and unidirectional arrows. The moral is that not every outline of unclear ideas, though somehow graphically presented, may be recognized as a path diagram.

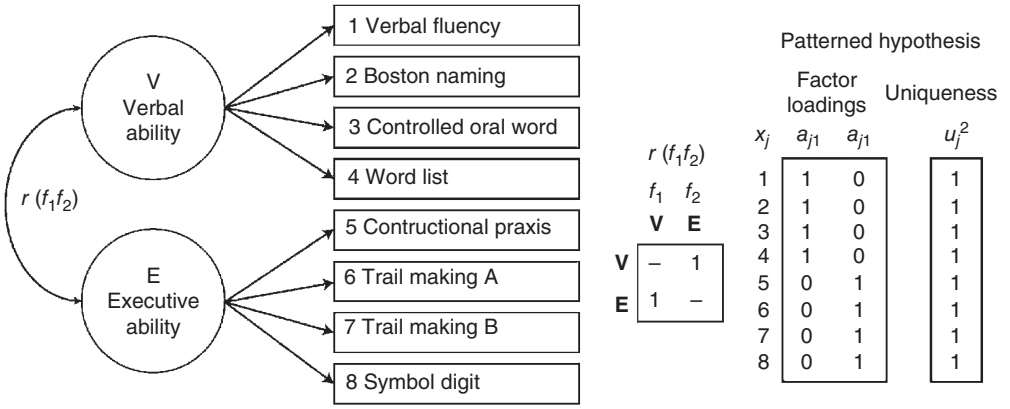


Figure 7.5 Path diagram and patterned matrices defining the confirmatory factor analysis of the two-factor ICS hypothesis to be tested for the correlation matrix of Table 7.1.

2) *Patterned matrices*, in simplest form, are expressed in terms of (0; 1) tables, which are in a one-to-one correspondence with the above-mentioned path diagrams. In this simplest form, their elements indicate an existence (by 1) and/or nonexistence (by 0) of regressions (→) in one matrix, and correlations (↔) in another matrix. Thus, after having abandoned the hypothesis of unidimensionality of the eight subtests of the CERAD in Table 7.2, we can formulate a new prescribed structural hypothesis of two independent clusters also using patterned matrices and a branching diagram as shown in Figure 7.5.

For the test of the two-factor ICS hypothesis outlined in Figure 7.5, we again ran a CFA using this same Stata *sem* procedure on this same 8×8 CERAD data matrix. The results are shown in Table 7.3. The fit of this two-factor ICS hypothesis is still bad—see the "unexplained" correlations larger than 0.30 in magnitude, and several over 0.20, in Table 7.3. Regardless of the fact that the overall fit index CFI is a little better than in the case of a single *g*-factor, the several high residual correlations indicate that the hypothesis of two independent clusters (though correlated) is not acceptable. It is important not to rely only on the one-valued global fit indices, which aggregate, and in a certain sense, "average" the misfit of the prescribed structural hypothesis. Therefore, although an overall fit may look reasonably good, there still can be several quite high "unexplained" correlations (e.g., greater than 0.3 absolute magnitude) remaining in the residual correlation matrix. In any case, *a visual check of all residual correlations should always be done.* In research memos and journal papers, the residual matrix should be presented along with the matrix of loadings, fit indices, etc., to present the evidence to the readers and to enable other researchers to verify the reproducibility of results.

7.2.2 (§67) Confirmatory vs. Exploratory Simple Structure

In Table 7.3, it is obvious that subtest 5, Constructional Praxis, does not agree with the hypothesis about its belonging to the second cluster, which was expected to represent Executive ability, *E*. The same was already suggested by the manifest structure in the initial correlation matrix of Table 7.1, where subtest 5 had low correlations with the other subtests of cluster *E*. Therefore, one of the possibilities to improve the fit may be to allow this indicator to become equivocal for both LC factors. However, it is difficult to find substantive support for doing this, which should be based on the semantic-content validity of the test task of subtest 5, and why it should also be recognized as

Table 7.3 Fitting the prescribed patterned hypothesis from Figure 7.5 using confirmatory factor analysis. The two-dimensional ICS structure of the correlated LCFs is still not good.

(Estimated with Stata sem, Maximum Likelihood method)												
			Patterned matrix of the modified structure						Output solution, computed values			
			x_j	$\boldsymbol{V}$	$\boldsymbol{E}$				f_1	f_2	u_j^2	
			1	1	0				0.69	0	0.53	
	$\boldsymbol{V}$	$\boldsymbol{E}$	2	1	0		f_1	f_2	0.68	0	0.53	
$\boldsymbol{V}$	—	1	3	1	0	f_1	—	0.29	0.68	0	0.54	
$\boldsymbol{E}$	1	—	4	1	0	f_2	0.29	—	0.72	0	0.48	
			5	0	1				0	0.11	0.99	
			6	0	1				0	0.97	0.05	
			7	0	1				0	0.91	0.17	
			8	0	1				0	0.44	0.81	
		Residual correlations										Fit
		1	2	3	4	5	6	7	8			
	1	—										$CFI = 0.87$
	2	0.03	—									$SRMR = 0.15$
	3	0.05	−0.06	—								
	4	−0.07	0.05	0.01	—							$\lvert r_{RES}\rvert max = 0.46$
	5	**0.36**	**0.46**	**0.31**	**0.38**	—						
	6	−0.01	−0.09	−0.01	0.06	−0.01	—					With **eight** additional high unexplained $\lvert r_{RES}\rvert$ values **over 0.10**
	7	0.02	−0.08	0.02	0.05	0.00	0.00	—				
	8	**0.13**	**0.11**	**0.19**	**0.29**	**0.18**	0.00	−0.02	—			

an indicator of verbal ability. Therefore, we must consider the following procedure only as a didactic illustrative example of the technical aspects of prescribing a modified hypothetical structure. Then the newly prescribed hypothesis takes on a form shown in Table 7.4—but the indices still indicate a bad fit.

Table 7.4 A confirmatory liberalization of ICS from Table 7.3 to a more general simple structure. However, even a non-ICS structure with the equivocal indicator 5 did not fit.

(Estimated with Mplus, Maximum Likelihood method.)												
	Patterned matrix of the modified structure							**Output solution, computed values**				
			x_j	$\boldsymbol{V}$	$\boldsymbol{E}$				f_1	f_2	u_j^2	
			1	1	0				0.68	0	0.54	
	$\boldsymbol{V}$	$\boldsymbol{E}$	2	1	0		f_1	f_2	0.71	0	0.49	
$\boldsymbol{V}$	—	1	3	1	0	f_1	—	0.28	0.65	0	0.57	
$\boldsymbol{E}$	1	—	4	1	0	f_2	0.28	—	0.72	0	0.47	
			5	1	1				0.60	−0.06	0.66	
			6	0	1				0	0.98	0.04	
			7	0	1				0	0.90	0.18	
			8	0	1				0	0.44	0.80	
		Residual correlations										**Fit**
		1	2	3	4	5	6	7	8			
	1	—										$CFI = 0.96$
	2	0.01	—									$SRMR = 0.08$
	3	0.08	−0.07	—								
	4	−0.06	0.02	0.03	—							$\lvert r_{RES}\rvert max = 0.29$
	5	−0.02	0.06	−0.05	−0.01	—						
	6	0.00	−0.10	0.00	0.07	0.00	—					With **four** additional unexplained $\lvert r_{RES}\rvert$ values over 0.**10**
	7	0.03	−0.08	0.03	0.06	0.00	0.00	—				
	8	*0.14*	*0.11*	*0.20*	**0.29**	*0.18*	0.00	−0.02	—			

7.2.2.1 Simple Structure

We use the illustrative example to distinguish the types of simple structure introduced in Section §64 more explicitly: (i) An *approximate* Simple Structure, defined in terms of an optimal mathematical criterion that is similar to the Varimax principle; this type cannot be fitted by a confirmatory approach and needs a kind of rotation and (ii) a nonapproximate simple structure,

i.e. *prescribed by a patterned matrix*, to fit it to the correlation matrix, whose very special case is the ICS, but also a somewhat more general case of *non-ICS* simple structure.

The hypothesis in Table 7.4 is an instance of the *confirmatory*, i.e. *prescribed non-ICS* simple structure. In fact, it is an instance nearest to the strict ICS since there is just one indicator which breaks ICS due to its equivocality. A little strangely, indicator 5 exhibits validity of 0.60 to the factor *V*, contrary to the former hypothesis. Its factor validity to the LC factor *E* which became a so-called secondary loading is almost 0, which means that this indicator is almost univocal but measures a quite different theoretical concept than originally expected. However, in a more general confirmatory simple structure, it may be intentionally presumed that both loadings on the two different LC factors are important, perhaps almost equal, possibly exhibiting a compensatory combination or "competition" of two abilities (as mentioned in Section §64). Of course, usually it is not assumed that there could be more than a few such nonunivocal indicators. Moreover, such a hypothesis should be based on substantive content and reasonable theoretical grounds.

On the contrary, an *exploratory approximate Simple Structure*, "discovered" in terms of an exploratory approach to the application of the LCF model, generally allows all loadings to be different from zero. In many cases, it happens that even the simple structure produced by a formally optimal rotation method, such as Varimax or Oblimin, still does not seem to be "very simple" at first sight. It is necessary to make a distinction between the *exploratory* version of the *Latent Common Factor* (LCF) *model* and the so-called "exploratory factor analysis" (EFA), especially when it is in the form of PCA. The latter so-called "factor analysis" is typically (and unfortunately) done by the computational methods of principal component analysis (PCA), or its version known as principal axes factoring. On the other hand the *exploratory LCF approach* can be applied in either one of two variants.

7.2.2.2 Exploratory Common Factor Analysis

The *fully exploratory* version, which is the first one, includes the exploratory search for the *unknown number m* of LC factors that might "explain" the manifest correlations. This case seems to be rather problematic as the researcher has almost no idea what he/she might expect to get, perhaps just to learn about the dimensionality of a collection of tests or items. It seems that such an approach does not represent a methodologically based variant of scientific inquiry. The only difference from a "blind factor analysis" by PCA is that this version is focused on finding the *dimensionality of the correlations without variances*. Hence, the artificial *f*-variables are genuine LCFs, i.e. "explaining" what the manifest variables have in *common in their inter-relationships*. In addition, this approach bears one of the important features, namely that—in theory—*it can fit the correlations perfectly in its parsimonious version* ($m < n$) while PCA is able to fit them perfectly only if the number of components (or "principal factors") is increased up to the number of original observed variables, i.e. only with $m = n$.

The second version is the *partly* exploratory and/or "*the weakest confimatory*" search for unknown structure but with a *prescribed number m* of LC factors. It seems to be a "minimally confirmatory" approach with a minimum restriction of the LCF model, as one at least has a *hypothesis about the dimensionality* of the inter-correlations. Even in the case where such an exploratory simple structure cannot allow the researcher to formulate any ex-post semantic interpretation, this procedure still accomplishes a certain formally plausible aspect of *a scientific model*—the number of LC factors represents a generally *falsifiable*, and empirically testable hypothesis, which may fit perfectly in theory. It can have a straightforward *practical use* in some special situations.

For example, Professor Blahuš once cooperated on a cross-cultural study whose aim was to compare a four-factor structure for a questionnaire across several different populations, but one sample dramatically differed from the other three samples in the semantic interpretation of the factors.

Hence, different modifications were tried on the last sample to find a structure, which would at least resemble the semantic interpretation in the other three samples. Several unsuccessful attempts with that sample were finished with the last minimal hope that the questionnaire might at least be four-dimensional regardless of any prescribed structure. However, it was not—even six unspecified exploratory factors were not enough. Probably, the cultural and environmental character of that sample caused a quite different perception of the questions by the respondents.

Similarly, we may free the "almost-ICS" hypothesis in Table 7.4 still further and try an approximate exploratory simple structure hypothesis limited minimally just by the number of factors, $m = 2$. For that purpose, we can use standard SPSS software, which also offers the Least Squares method for fitting LC factors and several methods for the approximation of general simple structure. The analysis of the correlation matrix in Table 7.1 (Section §65) yields the basic solution whose loadings do not yet exhibit features of a simple structure, that is, a clear and interpretable factor solution. Therefore, the mathematically optimal transformations to approximate simple structure for uncorrelated factors (Varimax) as well as for correlated factors (Oblimin) are presented in Table 7.5.

7.2.2.3 Factor Rotation

The exploratory fit of our dimensionality hypothesis $m = 2$ is evaluated on its basic solution. This fit is not changed by varying transformations/rotations. The index $GFI_{LS} = 0.996$ remains the same over the three solutions in Table 7.5. From this point of view, all three solutions are *mathematically equivalent* but offer *different semantic interpretations*. Figure 7.6 presents the graphical result of the Varimax rotation of axes. In other words, the x/y coordinates for the vectors for subtests 1 through 8 in Figure 7.6 are given in the Varimax loadings under (i) in Table 7.5. The vector orientation before rotation is given by the loadings as coordinates in the basic solution. Obviously, the configuration of points (indicators) is not changed by a rotation. If we wished, we could rotate the Varimax solution in 7.5 back to get the same loadings as in the basic solution or in the Oblimin solution.

To graph the vector configurations for various rotations, we would proceed in the same way as Thurstone about eighty years ago—by using graph paper and a ruler. Thurstone proceeded from the basic solution to the rotated version to find such an angle of rotation that would yield numerical values of the coordinates of the points (indicators) large along one axis and small along the other one. This is a *graphical formulation* of a general principle of approximate simple structure in its exploratory form. In practice, it is impossible to reach the exact ICS structure as the strongest variant of the general simple structure by additional rotation/transformation of the so-called basic exploratory solution. Transformation with correlated factors involves the change of the orthogonal axes (uncorrelated factors) into oblique axes (possibly correlated factors). Then the cosine of their angle is equal to the correlation between the factors. Imagine how tedious it had to be when Thurstone and his collaborators were rotating all the combinations of pairs of many factors (originally more than six), and each approximation to an intuitive assessment of the plausibility of a simple structure had to be repeated many times and improved step by step.

Regardless of the good fit of the two LC factors in Table 7.5 (as shown in the small residual correlations), there are too many equivocal indicators, and any reasonable semantic interpretation seems to be impossible. In the graphical configuration in Figure 7.6, we also can see a different clustering of indicators than was formerly hypothesized. Hence, although we have found that the correlation matrix of the eight tests is factorially two-dimensional, this finding is only a formal mathematical fact that need not be somehow related to some possibly substantive theoretical concepts, say abilities—compare the problem of dimensionality due to "difficulty factors." Now, these not very plausible findings will be used to illustrate some additional problems in the relationships between formal dimensionality and semantic homogeneity.

Table 7.5 Switch to an *exploratory model* (EFA) for two LC factors for a rotational approximation to simple structure, with the hypothesis that "two dimensions may explain the correlations."

	Initial solution		**Simple structures approximated by rotation**											
	Unrotated		**(i) Varimax uncorrelated factors**		**(ii) Oblimin correlated factors**		**Residual correlation matrix**							
	1	2	1	2	1	2								
1. Verbal fluency	0.61	0.32	0.67	−0.14	0.68	−0.25	—							
2. Boston naming	0.58	0.44	0.72	−0.03	0.75	0.10	0.00	—						
3. Contr. oral word	0.55	0.24	0.58	−0.17	0.58	−0.07	0.09	−0.07	—					
4. Word list	0.69	0.28	0.71	−0.22	0.70	−0.11	−0.03	0.03	0.00	—				
5. Constr. praxis	0.49	0.37	0.61	−0.03	0.63	0.08	−0.03	0.01	−0.02	−0.03	—			
6. Trails making A	−0.69	0.72	−0.08	0.99	0.07	**1.01**[a]	0.00	−0.02	0.03	0.01	−0.01	—		
7. Trails making B	−0.63	0.63	−0.09	0.89	0.04	0.90	−0.02	−0.01	−0.01	0.02	0.01	0.01	—	
8. Symbol digit	0.51	−0.11	0.32	−0.41	0.28	−0.37	−0.05	−0.05	0.03	0.06	0.02	0.00	0.01	—
					$r_{f1f2} = 0.31$									
							$\lvert r_{RES}\rvert max = 0.09$							

[a] Not a mistake: With correlated factors, the loadings become just regression coefficients $a_j = \beta_j$, and may exceed |1|, cf. Section §64.

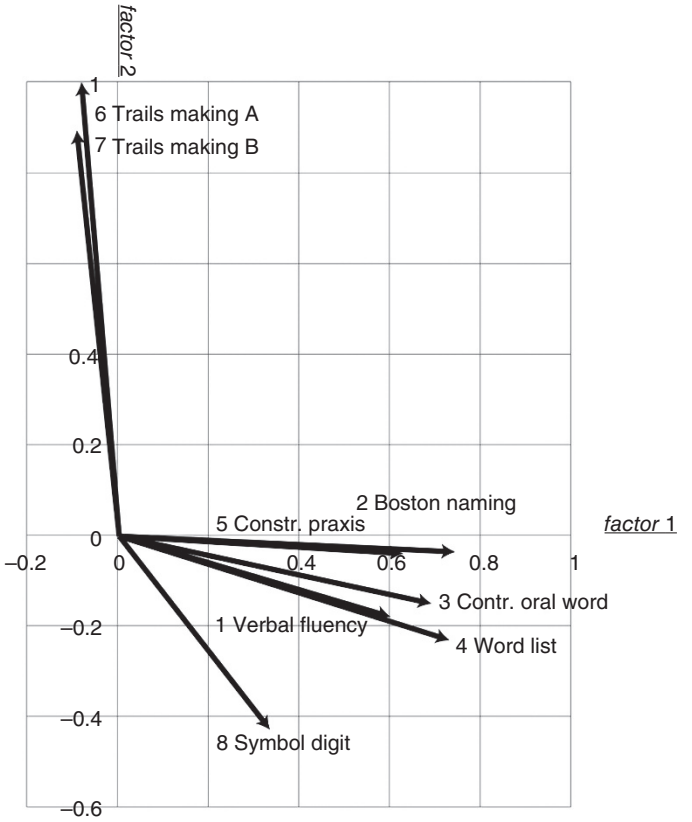


Figure 7.6 Graphical representation of the indicators in the two-dimensional LC factors *exploratory model* after the Varimax rotation yielding uncorrelated, orthogonal factors. The *x*–*y* coordinates are from the loadings under (i) in Table 7.5.

7.2.3 (§68) Stronger vs. Weaker Support for Content Homogeneity: By Unidimensionality or ICS

Unidimensionality of inter-correlations (without variances) as a syntactic, mathematically formalized counterpart of assessing the *content homogeneity* of a "construct" is most transparent in a *confirmatory* approach, especially with a *prescribed ICS*. It means that each cluster of indicators is intentionally selected by a hypothesis to be of the same semantic content, and exclusively of this content. Nevertheless, this situation may be incorporated into different circumstances and combined with various aspects.

A situation may occur when a one-layer structural hypothesis seems to be so plausible that the researcher dares to prescribe the ICS as a whole and fit it in one step. Besides such a *one-stage* complete ICS fitting procedure (as in Tables 7.3 and 7.4), the structure can also be verified in *two stages*:

i) Fitting a single (general) factor common to each of the presumed clusters *separately*, and when these separate unidimensional structures are acceptable, then
ii) Fitting these several "separately general" LC factors *simultaneously as a prescribed Simple Structure* at once.

Our unsuccessful confirmatory attempt in Table 7.3 is actually rather close to stage (ii). Namely, for the first cluster, verbal ability, stage (i) has been plausibly completed already in chapter 3, Table 3.2, section §33a and was repeated again in Chapter 6, Table 6.1, section §57a.

The variants (i) and (ii) are not just different versions of a data processing procedure, they relate to a more general methodological question of whether, and under what conditions, unidimensionality may be a good model of semantic homogeneity. It seems obvious that the *stronger* case of unidimensionality as a model of content homogeneity is found if there is a good fit in *both* of the above-presented stages (1) and (2). It means each cluster in stage (1) is separately unidimensional, and the multidimensional prescribed simple structure in stage (2) is plausibly fitted.

7.2.3.1 Higher-Order Factor Analysis

However, fitting a *non-ICS* structure at once as a whole—either prescribed (as in Table 7.4) or just exploratory (as in Table 7.5)—is accompanied by two circumstances:

a) It is usually based on the assumption that the "constructs" are not mutually independent in the corresponding theory; hence the model is typically chosen with LC factors that are mutually correlated. The correlations between the first-order factors might be interpreted as if that part of a univocal indicator, which it has "secretly in common" with univocal indicators of another factor, were transferred and hidden into the inter-factor correlations.

   According to older procedures for higher-order analysis, the correlations between these first-order LC factors can be analyzed to fit a possible unidimensional second-order general factor. If the fit were plausible, it would support the above-mentioned "hidden commonality behind the scenes" of the former one-layer ICS solution with separate unidimensionalities of its first-order factors. In the case of the recently more usual simultaneous fitting of the two-layer hierarchical structure as a whole, this "commonality" of the first-order ICS factors will be projected into the second-order layer of hierarchy. It would influence the second-order loadings of the first-order factors on the second-order factor, and the overall fit of the model as a whole. Although the structure was of the ICS type, the inter-factor correlations indicate that there is a kind of an "overlap at a higher layer," which seems not to follow a strict exclusivity of the indicators. Nevertheless, this seems to be a rational necessity in theories with a hierarchical structure—for decades, there were discussions about the structure of intelligence between the advocates of its sectioned character in Thurstone's sense and those who were still holding to the idea of its holistic character in the sense of Spearman's *g* at least on a rather high layer of hierarchy.

b) A more important aspect is related to the equivocal indicators. For instance, if we fitted a simple structure with two LC factors, the resulting structure may have the form of an "almost ICS" similar to the prescribed patterned matrix in Table 7.4, which has only one indicator with complexity 2. Then, if we intend to use the strategy of fitting single (general) factors to the individual clusters separately, we face the problem of deciding which cluster the equivocal indicator should preferably be included for the test of unidimensionality. We may include the equivocal indicator only in that cluster to whose factor the indicator had higher factor validity, its former "primary loading."

In our example, the results in Tables 7.4 and 7.5 show that test 5, constructional praxis, has a higher, primary factor validity to the verbal cluster, which is quite opposite to the original hypothesis, and there is no simple semantic interpretation of this finding. The situation may suggest checking the hypothesis also by an attempt to fit a single factor to the separate cluster hypothesized as executive ability with test 5 as "secondary"—as was indicated in stage (1) above. This separate analysis of the 4×4 correlation submatrix of the executive cluster in Table 7.1 has shown that test 5 does

not follow the assumption of unidimensionality with the other three tests. From the limited point of view of this illustrative example of our eight tests, the situation asks for a revision of the original structural hypothesis. It may lead either to a more fundamental revision of the role of the test in the whole battery or just to a simple elimination of such a "strange" subtest—see Table 7.6.

7.2.3.2 Equivocal Indicators and Factorial Complexity

In general, and quite theoretically, we may consider an equivocal indicator that has a higher (primary) factor validity in the frame of a certain cluster C1, which is separately unidimensional, but the indicator also has a secondary validity in the frame of another separately unidimensional cluster, C2. Say that the semantic content of cluster C2 is basically different from C1—as indicated by its own typical (primary) indicators. However, this still does not eliminate the possibility that C2 might stay unidimensional even if some further indicators of C1—formerly primary indicators of C1 and secondarily of C2—are now included in C2. Hence, an equivocal indicator might be included in more than one cluster while each of the clusters may be still found separately unidimensional. Moreover, some such indicators might be almost literally "*equi*-vocal" having almost the same factor validity as each of the two LC factors, say if both of the two abilities were equally "important" for a good performance in the indicator's task.

In addition, it is quite likely that the unidimensional factor validity of the formerly secondary indicators will be higher due to the absence of the other "competing" LC factor. The omitted factor now does not attract the other part of the true score of the indicator, and this part will likely be included in the unique part, and its variance into the new uniqueness (u_j^2). It is necessary so that any indicator with *prescribed complexity two* (or more) was given appropriate theoretically based evidence of its *complex content*. A detailed substantive analysis should be done to explain why and how the indicator was intentionally constructed to represent two abilities or traits simultaneously.

Even in the case of a prescribed ICS, certain parts of indicators' variances are projected into the correlations between the LC factors. Nevertheless, by a comparison of ICS with the same situation in the non-ICS case, though each had separately unidimensional clusters, ICS supports content homogeneity the *strongest* due to indicators' *exclusivity*. On the other hand, we also may have further *weaker versions* of unidimensionality supporting content homogeneity:

a) A similar one to the *strongest version but fitted simultaneously with a higher-order factor* (or factors), which means that indicators' exclusivity on the first-order layer has enabled them due to their hidden involvement in the higher-order layer common relationships. This situation is often found in a priori hypothesized multilayer theories of a more complicated character. These usually assume an ICS structure of indicators on the lowest layer but with some common higher-order factors—see the examples of health-related fitness in Figure 7.2.
b) The separate clusters are unidimensional but in the frame of a simultaneous multifactor simple structure, the overall fit needs to prescribe a non-ICS-type structure with some more complex indicators. This is a situation of "unidimensionality without exclusivity." Its special case is represented by a bi-factor structure where the group factors resemble ICS aside from the *g*-factor (Figure 7.1a,b).

7.2.4 (§69) "Hidden Equivocality" in ICS Structures and the So-Called Frame Problem

At first glance, and by tradition, the prescribed ICS structure in the confirmatory approach may seem to support the hypothesis about content homogeneity and "construct" separability very well.

Table 7.6 Confirmatory LCF model after the elimination of test 5, a revision of the ICS structures from Tables 7.3 and 7.4 (Estimated with Stata sem, Maximum Likelihood method).

	Patterned hypothesis		**Estimated loadings**		**Residual correlations**							**Fit indices**
	V	***E***	**1**	**2**	**1**	**2**	**3**	**4**	**6**	**7**	**8**	
1. Verbal fluency	1	0	0.69	0	—							
2. Boston naming	1	0	0.68	0	0.03	—						$CFI = 0.95$
3. Contr. oral word	1	0	0.68	0	0.05	−0.06	—					$SRMR = 0.08$
4. Word list	1	0	0.72	0	−0.07	0.05	0.01	—				
6. Trails making A	0	1	0	0.98	0.00	−0.09	−0.01	0.07	—			
7. Trails making B	0	1	0	0.91	0.02	−0.08	0.02	0.05	0.00	—		Four residuals over $\|0.10\|$
8. Symbol digit	0	1	0	0.44	*0.13*	*0.11*	*0.19*	*0.29*	0.00	−0.02	—	
			$r_{f1f2} = 0.29$									

However, we saw above that there still might be something common to the indicators from different ICS clusters. This "hidden commonality" (not communality!) can take on a form of quite highly correlated first-order LC factors, or the possibility of fitting higher-order factors. In addition, there is another related problem connected with the fitting of separate unidimensionality mentioned in the last section. It may be called "hidden equivocality."

Assume that in a study a cluster C1 fits a single-factor model with its general LC factor *f*, on which an indicator has a quite important factor validity, such as the loading of 0.60 in Table 7.4 (Section §67). Further, imagine that another cluster C2 of indicators is not included in the given research project or, even, that the existence of the further "construct" 2 and potential cluster C2 is completely unknown in the given context. Then, even a quite high factor validity (such as the mentioned 0.60) of the indicator in the frame of the former unidimensional cluster C1 does not mean that it is exclusively measuring only "construct" 1 modeled by *f*. The indicator might still measure a further "construct" 2 with equal or even higher single-factor validity when incorporated into cluster C2 alone in a separate confirmatory unidimensional LCF model of C2 if we had the indicators creating C2 at our disposal.

In addition, we could also try to fit a confirmatory two-dimensional "almost ICS" model similar to Table 7.4, allowing this one "strange" indicator to have nonzero loadings on both LC factors. In this new situation with the set of indicators consisting of both clusters C1 and C2, hence with two factors say f_1 and f_2, let the former factor *f* of the original insulated cluster C1 become the new factor f_1. Now, we may prescribe the pattern of loadings so that the "strange" indicator had its loading on the new f_1 equal to the loading it had on the former *f* (say 0.60) and wait for its resulting loading on f_2 after the computational fitting of both LC factors. Then, mathematically, the indicator with a loading of 0.60 on f_1 still may have a high loading, say 0.70, on f_2. This is due to the fact that (e.g. with uncorrelated factors) the indicator's hypothetical communality of the joint set of C1 and C2 indicators would be $0.60^2 + 0.70^2 = 0.85$, still leaving 0.15 for the indicator's uniqueness (cf. Section §68). However, since we might not even know about the existence of "construct" 2, we are not in a position to create the second battery—cluster C2—of indicators measuring the second factor f_2 and carry out the research study with two factors.

7.2.4.1 Indicator Multidimensionality

It may seem that what is considered a specific component of a certain indicator in the context of one content-homogeneous cluster might be composed of the indicator's validity to infinitely many other clusters. In addition, this also seems to oppose the idea of a homogeneous domain/subdomain assumed by generalizability theory in the sense of the sometimes-claimed opinion that "No test measures a single factor ... a test which is truly 'univocal'" (Vernon, 1971, p. 133). Moreover, the indicator might often be found to be unidimensional when incorporated separately into each of such different clusters. From the other point of view, indicators in many other unknown clusters might leave our unidimensional cluster still more unidimensional if incorporated into it. Of course, the clusters represent samples of tests/items from infinite different domains or subdomains, and the cluster at hand is always only a finite and very limited sample. Therefore, we must have a very well-based, especially theoretically based, substantive semantic meaning of the domain and the content validity of each of its indicators.

Otherwise, we may face the problem of motivating a question such as: Yes, I think that this test measures this "construct" but perhaps it also measures some other "constructs," how could I know? How could I practically make a complete list of them? If we do not have a theory we have to admit that there is an infinite number of "constructs," and that there is also an infinite number of different variables and situational or contextual conditions that might influence our observations/testings, experiments, etc. This would lead us into pure agnosticism and make scientific inquiry completely

meaningless. If we have no theory, we are not able to involve the conceptually important "constructs" and domains in consideration. Hence, any research would seem useless—even if we would like to start, we do not know which "construct," which indicator, and which contextual condition we should start with.

Actually, in some situations where it is really important to have a lavishly concrete (maybe an almost complete) list of all instances of "constructs" and indicators, a researcher gets into a difficult position. A complex study of "wellness" can be an example. However, this dangerous situation may occur if any kind of complex research project is supposed to cover a very wide area consisting of several domains, say of sociological, psychological, biomedical, and other aspects such as in some epidemiological surveys or studies that bear a character of "data mining."

7.2.4.2 The "Frame Problem"

Such a situation is related to the so-called "frame problem," a well-known problem in Artificial Intelligence—AI (see e.g. Dennett, 1987). In the discussed circumstances, the frame problem deals with the difficulty or impossibility of making a definite list of all possible "constructs"/LC factors, to which the given indicator might have a semantically meaningful non-negligible "construct" validity/factor validity. Or, if such a "*super-domain*" of all different domains of the separate "constructs" were at hand, then the difficulty would consist of *selecting the relevant* ones. Moreover, if those that are relevant were known and selected, it would be still impossible to carry out their joint research study and analysis in practice due to their huge amount.

From the opposite point of view, this problem shows the importance of the hypothetical idea of an infinite domain of indicators that should be imagined by a researcher during the process of test construction. It also holds for constituting any set of diagnostic tools, including biomedical laboratory measurements, questionnaires, inventories, etc. Otherwise, the researcher is in danger of being unaware that his/her approach is based on a simpleminded reduction to several separately selected variables. In general, awareness of the frame problem suggests a recommendation for research practice, and it reminds us of one of the fundamental principles of scientific inquiry: namely, the importance of the conceptual, *theoretical ladenness* of our hypotheses that are to be verified/corroborated by empirical research. In the current context, we must start with a theoretically well-founded hypothesis on the structure of a domain. Of course, there are usually several competing hypotheses but each of them must be thoroughly based on a preceding theoretical account as well as empirical evidence. Of course, each of the theories as well as empirical research projects are based on a *methodological simplification*.

The frame problem also has an additional important methodological consequence. Namely, most researchers tend to form hypotheses as selected "positive" statements, i.e. positively related to the solution of their research problem. This is an approach acting as if these statements were drawn from a pool where all other statements, conditions, and variables were indifferent to the problem. For instance, a causal hypothesis may state that some experimentally manipulated variables are supposed to influence the outcome variable. Thus, all other possible variables seemed to be left in the infinite pool of temporarily indifferent variables from which some may be selected as causal ones and say added in the next experiment.

However, awareness of the frame problem should teach us that the conceptual-theoretical formulation of a hypothesis should also include "negative" statements. We should bring theoretical evidence and clearly say as a part of our hypothesis which of the variables are *not* supposed to be possible causes, i.e. that this and this variable are *unimportant*. In the structural studies on "construct"/factor validity, the frame problem suggests a strong appeal to bring some *evidence of irrelevance* and to also present "negative" counterparts of the structural hypothesis. It means that it should be clearly stated which of the possible and sometimes considered indicators are excluded

from the study, which of the "constructs" are not important in the given theory, etc. In a sense, it also corresponds to the system analytic approach which draws a line between the objects that belong to a set and those that do not. For example, in the current context, it is also necessary to say which indicators (or what type of indicators) were evaluated, and why they are or are not considered to represent a domain of certain "constructs," etc. Different variants of a scientific theory composed of theoretical concepts of various generality, and organized on different layers of hierarchy, should be conceptualized with stress on a detailed characterization of the nature of the concepts from the above-indicated points of view.

7.2.5 (§70) Dimensionality by the LC Factor *Model* vs. PCA "Factor *Analysis*"—Example

If factor unidimensionality (as opposed to multidimensionality) is to be considered a formal model of semantic-content homogeneity (as opposed to heterogeneity), then it is necessary that certain minimum conditions of such mirroring are met. Although we are going to comment on the general topic of modeling in science later, let us point out that one of the necessary conditions is a general requirement of falsifiability of a model based on a substantively appropriate test of model fit. As was mentioned several times, the unclear term "factor analysis" very often misleads an uninformed user into using certain methods that do not meet these minimum conditions. Typically, one of these methods, which some people consider to be "The Factor Analysis," is PCA—principal component analysis (and/or its iterative variant called "principal axes factoring"), as was mentioned above. Besides the lack of the minimum properties of PCA to work as a genuine scientific model, I will now present an example to show how *PCA is inappropriate for verification of the dimensionality of* "*constructs*," or for verification of the *substantive hypothesis about content-homogeneity* of a cluster of indicators chosen to measure their common "constructs."

7.2.5.1 Explaining Correlations vs. Explaining Variance-1

It can be seen below that while the LCF model "explains" the mutual intercorrelations of the indicators almost perfectly by fitting only one factor (i.e. with $m = 1$ dimension), the PCA "factor analysis" will need three principal components/"factors" to approach a comparable result (i.e. with $m = 3$ dimensions). The analyzed data are former correlations of the four verbal ability tests. Thus, the LCF model strongly supports the concept of verbal ability as a content-homogeneous attribute almost perfectly "explaining" the *common relationships—intercorrelations* among the verbal subtests. On the contrary, the corresponding application of PCA offers false evidence that verbal ability is a heterogeneous multidimensional concept. This is due to PCA's mathematical focus on the *individual indicators' variability*—i.e. on the indicators' *variances* but *not* on their mutual correlations. In Table 7.7 (a), we have placed the residual correlation matrix that results from analyzing the first four subtests of the CERAD battery using effective LCF models with $m = 1$ factor. This matrix has been imported to this chapter from Table 6.1 of Chapter 6 (section §57a) for a direct contrast of a strong and effective LCF analysis as compared with the weakness of the residual correlation matrix (b), from an analysis of the same matrix with $m = 1$ factor using PCA.

Sole, insulated use of PCA yielding such high residuals in Table 7.7 (b) would offer a misleading result perhaps indicating that the researcher should try PCA with $m = 2$ components. The researcher, hardly having a concrete alternative conceptual hypothesis about a more complicated structure of verbal ability represented only by the four tests, may decide to use PCA with Varimax rotation to see if there could be a kind of exploratory indication for a certain simple structure. Then, a usual variant of data processing would typically (perhaps in SPSS) offer an output mostly consisting of the results in Table 7.8.

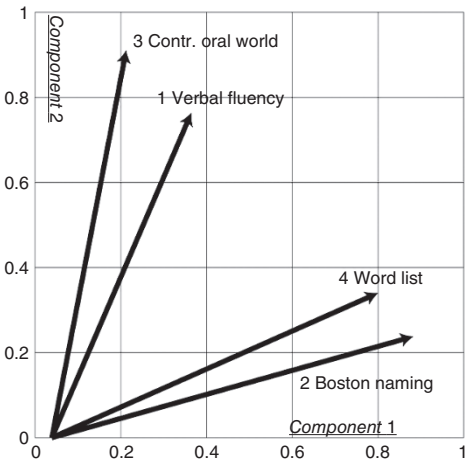
Table 7.7 Residual correlations of the unidimensional $m = 1$ latent common factor (LCF) model from table 6.1 vs. PCA with $m = 1$ component.

	Original correlation matrix				Latent common factor model (a)					Principal component analysis (b)				
					Factor loadings	Residual correlations				Component coefficients	Residual correlations			
x_1 V-Flue	1.00	0.51	0.52	0.43	0.70		0.02	0.04	−0.07	0.79		−0.11	−0.08	−0.20
x_2 B-Nam	0.51	1.00	0.40	0.54	0.70	0.02		−0.09	0.03	0.79	−0.11		−0.20	−0.08
x_3 COWo	0.52	0.40	1.00	0.50	0.69	0.04	−0.09		0.01	0.77	−0.08	−0.20		−0.11
x_4 WordL	0.43	0.54	0.50	1.00	0.71	−0.07	0.03	0.01		0.79	−0.20	−0.08	−0.11	

Table 7.8 A dubious, but often found, "Exploratory FA by PCA with Varimax" with $m = 2$ components.

	Residual correlations				Component matrix of the component coefficients: Unrotated component 1	Unrotated component 2	Rotated component 1	Rotated component 2
x_1 Verb Fl		0.02	−.22	−0.10	0.79	0.29	0.37	**0.75**
x_2 B Nam	0.02		.01	−0.22	0.79	−0.44	**0.87**	0.23
x_3 CO Wo	−0.22	0.01		0.04	0.77	0.49	0.21	**0.89**
x_4 Word L	−0.10	−0.22	.04		0.79	−0.32	**0.79**	0.32

Varimax rotated "Loadings Plot"



The results of the approximate simple structure found by Varimax rotation might suggest that verbal ability consists of two sub-"constructs," each of them measured by two tests. Post hoc exploratory considerations may lead the researcher to search for substantive reasons why tests 1 and 3 have higher coefficients on Component 2 while tests 2 and 4 on Component 1—see also the graphical plot. Perhaps, several creative and competing speculations might be considered, but hardly related to a more fundamentally based theory. Nevertheless, before starting to search for a possible semantic interpretation, the researcher should have a look at the residuals after the "extraction" of $m = 2$ components: Even the two components have "explained" the correlations very badly and the researcher might conclude that perhaps $m = 3$ components are necessary. Finally, with three components, the residuals in Table 7.9 (a) seem to be almost acceptably low, but what could be a substantive interpretation of the three would-like-to-be "constructs"? What might be a sense of such an approach or its scientific contribution in general, to use three dimensions to "explain" correlations of four variables, but still not too well?

7.2.5.2 Meaninglessness of "Communalities" in PCA-2

A more detailed understanding of the conceptual difference between the LCF model and PCA may come from the practical example in Table 7.9. Recall from section §56 that by the definition of the LCF model, a communality h_j^2 of manifest variable x_j is that part of its variance s_j^2 which is shared with the other manifest variables *through* their common LV f, which "explains" the manifest correlations—in the sense of partial uncorrelatedness. According to this definition, these *parts of the variances* s_j^2 which were "explained" by PCA should *not* be called "communalities," as these parts are *not* variances of the *common part*s $g_j = a_j f$ of the variables x_j (see Sections §68 and §69). Unfortunately, usual standard statistical software incorrectly prints them out under this title (keeping the spirit of the rough numerical methods before WW II).

Thus, on one hand, principal components reproduce incomplete parts of the s_j^2, while these parts are larger than necessary to "explain" correlations—see the comparison for $m = 1$ of the LCF model vs. PCA in Table 7.9. On the other hand, these PCA-generated parts, although larger, still "explain" less of the correlations, and contain some irrelevant variance of the variables, namely their undesirable specific, and unreliable error variance, which fact decreases the possibility to reproduce the correlations better. This happens due to a built-in mathematical property of PCA to reproduce primarily the scores and variances of the original variables x_j as far as possible, while correlations are reproduced as secondary consequents of these partly generated scores of x_j variables.

Another specific difference deals with the "percents 'explained'," although it is of a rather technical mathematical nature, it is quite often misunderstood in interpretations. If the LCF model were fitted by the method of least squares to the correlations of the four verbal tests, then the fit index $GFI_{LS} = 0.995$ means that an "overall inter-correlatedness" has been "explained" up to 99.5%, as measured by the sum of squares of the correlations, and the *off-diagonal* elements in the correlation matrix (see Section §66). On the contrary, the usual software output of PCA speaks about the "total variance 'explained'" (in terms of the eigenvalues of the correlation matrix), but without an explanation of what the "total variance" means, and why this should be used.

7.2.5.3 What "Variance" Is Explained by PCA-3

One may ask: "Explained" total variance of what variable? The answer is: If you make a sum of four z-standardized verbal test scores for each subject you will get a new variable, the simple-sum composite score, which will have its own certain variance—this is the "total variance" PCA deals with. (It follows from matrix algebra: The sum of the diagonal of a matrix equals the sum of its

Table 7.9 Four "unexplained," residual correlation matrices—LCF model vs. PCA.

					"Unexplained" parts of variances in the diagonals (typically not shown by the usual software)											
	LCF, *m* = 1				**PCA, *m* = 1**				**PCA, *m* = 2**				**PCA, *m* = 3**			
x_1 Verbal fluency	0	0.02	0.04	−0.07	**0.38**	−0.11	−0.08	−0.20	**0.30**	0.02	−0.22	−0.10	**0.09**	−0.10	−0.08	0.09
x_2 Boston naming	0.02	0	−0.09	0.03	−0.11	**0.38**	−0.20	−0.08	0.02	**0.18**	0.01	−-.22	−0.10	**0.11**	0.09	−0.10
x_3 Cont. oral word	0.04	−0.09	0	0.01	−0.08	−0.20	**0.41**	−0.11	−0.22	0.01	**0.17**	0.04	−-.08	0.09	**0.08**	−0.09
x_4 Word list	0.07	0.03	0.01	0	−0.20	−0.08	−0.11	**0.38**	−0.10	−0.22	0.04	**0.28**	0.09	−0.10	−0.09	**0.09**
	0.49	0.49	0.45	0.50	0.62	0.62	0.59	0.62	0.70	0.82	0.83	0.72	0.91	0.89	0.92	0.91
	Communalities h_j^2															
	Parts of variances that "explain" correlations				Parts of "explained" variances, **incorrectly** called "*communalities*" (by usual software)											
	99.5%				61%				77%				91%			
	Squared correlations "explained" by 1 LC factor (GFI_{LS} = 0.995)				"total" variance "explained" by one component				"total" variance " explained" by two components				"total" variance "explained" by three components			

eigenvalues.) This variance is then "explained" by the sequence of percentage parts given by the decreasing eigenvalues of the correlation matrix. It is quite unclear why one should be interested in "explaining" the variance of a simple sum of the original variables. Should one pay an interest in such an equally weighted, strangely mixed variable—especially if PCA were applied to the covariance matrix? The only exception seems to be a content homogeneous battery of z-standardized tests, but it is still a problematic question whether its composite score should be just a simple, unweighted sum. If there were two or more subclusters of tests in a larger battery, say one cluster for verbal ability, another for memory, ..., it seems highly questionable whether there exists a real-world interpretation of such a simple-sum score mixing together of different tests of different abilities.

Hence, also from a substantive interpretation, the "explained" percents of "total" variance is a dubious number. These percents, printed out by usual exploratory principal components "factor analysis" software at the bottom of Table 7.9, are percents of that variance of the simple sum variable: (i) principal component "explains" just 61% of that summed mixture, (ii) components "explain" 77%, and (iii) components "explain" 91%. Finally, the number of principal components equal to the number of variables, $m = n = 4$ would "explain" 100%. However, it is mathematically impossible that the so-called "incomplete PCA" with $m < n$ could fit any correlation matrix perfectly. From this point of view, PCA breaks another of the fundamental requirements of a scientific model, namely that in theory, *conceptually*, a model should be able to fit and generate data perfectly. All these circumstances eliminate PCA from using it for verification of semantic-content homogeneity of indicators as a hypothetical model.

As regards the mutual relation of syntactic dimensionality and semantic-content homogeneity, we should notice that only the so-called "complete PCA" with dimensionality $m = n$, i.e. with as many components as original variables, can reproduce the original correlation matrix perfectly. Then, it yields zero residuals also in the diagonal, which contains the variances—then it is not the case of the "incomplete PCA" with $m < n$ in Table 7.9. On the contrary, in the case of the LCF model, the variances in the diagonal must be perfectly reproduced by definition of the model since communality h_j^2 of each indicator is complemented by the indicator's uniqueness u_j^2 (which contains the irrelevant specificity and undesirable unreliability for reproducing correlations).

According to the common-sense meaning of the notion "unique," this property of the LCF model follows a principle that the *uniquenesses* cannot contribute to the identification of the *common* artificial variable f, which is intentionally constructed to "explain" what the indicators share jointly in their mutual relationships. From this point of view of the substantive interpretation of common factors, there is a built-in contradiction in the use of PCA. Since then, we have been using something unique to construct something common. This common f should be recognized as a formalized model of a theoretical concept in a scientific theory—and fundamental nomologic concepts of science should not be a priori conceptualized as intentionally constructed to contain specificities of different methods and their measurement errors.

7.2.5.4 Deterministic Component Scores vs. Factor Score Indeterminacy

There is still another aspect which also substantially distinguishes the LCF *Model* of a correlation/covariance matrix from the PC *Analysis* (including so-called principal axes factoring) as algebraic-numeric processing of any set of (real) numbers organized into a symmetric square matrix. The difference consists in the statistical character of the *LC factor scores* vs. the deterministic character

of the PC "scores" of principal components (incorrectly also titled "factor scores," in most commercial software). As we know from section §55, the true, exact LC factor scores of people on the LV *f* are unknown, we may obtain only their *regression estimates*, say by $\hat{f}_{BLS} = \Sigma b_j x_j$. On the contrary, the "scores" of a principal component are manifest *exact* linear combinations of the original manifest variables. This difference is also connected to the problem of LC factor indeterminacy (Section §70). Furthermore, it is also related to the aspects of the definition of latent vs. manifest variables in the strictly statistical, mathematical meaning.

However, in the present context, it is most important that the deterministic character of PCA "scores" makes them specifically given by the set of manifest variables at hand—by adding just one x_j, you will get quite *different* "*scores*," and it will happen every time you make a change, and every time the "scores" will be combined exactly but every time *differently*. Thus, such a process cannot be used for sequential improvement in identifying a "construct" formalized by the latent *f*. Namely, here, we miss the conceptualization of a hypothetically infinite content-homogeneous domain from which we may select more and more indicators to sequentially improve the regression estimates $\hat{f}_{\mathrm{BLS}}$ converging to the true *f* which is *just one*, unique. Hence, PCA cannot be used in this frame for conceptualizing *f*. Also, from this point of view, PCA cannot be used for the identification of "constructs" in the sense of genuine *theoretical concepts*.

Although PCA might be a fruitful method for descriptive data reduction or treatment of collinearity by orthogonalization, PCA is based on a different mathematical foundation than the LCF model of factor analysis, It is based upon a nonstatistical descriptive foundation. Its application follows different methodological principles: The principal component factors analysis is *not any model*, in a general meaning nor in a statistical meaning as was pointed out in Section §70. *Its sometimes roughly similar results, although seemingly reminiscent of the genuine LCF models, would be misleading*. We must say in conclusion to this section that *PCA "factor analysis" is inappropriate for the identification of theoretical concepts (constructs) and the verification or falsification of structural theories*.

7.3 Summary

Following Petr, we have moved not quite seamlessly between EFA and CFA with an emphasis on the common objectives and features of these two methods. Specifically, we used the correlation residuals to assess and compare the fit of different EFA and CFA models having two LVs. However, we did not directly address some key differences in the way EFA and CFA are usually applied. Whereas EFA is typically carried out with variables in standardized form as summarized by their (Pearson) correlation matrix, CFA is commonly conducted with unstandardized variables as summarized in their covariance matrix. Most CFA software provides optional postestimation of standardized model parameters (i.e. factor loadings) to correspond with the natively estimated unstandardized ones.

CFA software generally provides output of the covariance residuals, but not the correlation residuals. In principle, these could be derived mathematically from the original and model-implied covariances. The correlation residuals are available from certain specialized structural equation modeling (SEM) procedures that can be used to directly analyze correlation structures as well as covariance structures (i.e. EQS, SEPATH in Statistica, RAMONA in Systat). In Mplus, it is also possible to use a weighted least squares (WLS) estimator to directly analyze a

correlation matrix using individual participant data. With sufficiently large samples, these methods give unbiased estimates of standard errors and model fit significance tests and are preferred for general correlation structure analyses.

Standard errors of model parameters and statistical tests of model fit can be adversely affected when models that are not "scale invariant" are estimated by analyzing the correlation matrix as though it were a covariance matrix (Cudeck, 1989; Bentler, 2007). When models are scale invariant, model fit and related aspects of the model can often be appropriately assessed through direct analysis of the correlation matrix. Although determining whether a model is scale invariant can be complicated, Beretvas and Furlow (2006) provide an accessible account of scale invariance, indicating that many, but not all, factor models are scale invariant. Following Cudeck, they point out that a factor model will generally be scale invariant if the indicator residual variances are freely estimated, the factor variances are constrained to 1.0, and all factor loadings are either freely estimated or set to zero, without additional constraints beyond model identification. The models we were evaluating seem to satisfy these basic criteria, and the analysis of standardized variables was deemed to be usable and acceptable for our purposes in comparing the selected EFA and CFA model specifications.

In the process of finalizing the book, we did not have access to the original raw data for the EFA and CFA models, indeed, only the correlation matrix was available. Both Stata and Mplus (and various other SEM programs) can analyze summary data in the form of correlations, as we have shown with Stata in the computational exercises at the end of Chapter 7. When standard deviations are available, they can also be incorporated to generate the corresponding covariance matrix for analysis. When standard deviations are not supplied, both Stata and Mplus presume standard deviations of 1.0 (and means of 0) for the variables. This is equivalent to directly analyzing variables standardized to a mean of zero and a standard deviation of one.

7.4 Study Questions

7.4.1 A. Discussion Questions

7.1 Modeling Multiple and Hierarchical Constructs

§63 Multi-"construct" Theories—Illustrations by Some Two-Layer Structures

Q1 What do the authors mean when they say that "Holzinger's bi-factor theory was not a genuinely hierarchical theory"?

Q2 What are hierarchical factor analysis models? In what ways can they be useful?

Q3 What factor theory would they say is indeed a genuinely hierarchical theory and where is it diagrammed in Figure 7.1?

Q4 Compare Spearman's LCF "general intelligence" model to Thurstone's competing factor-analysis-based model of primary mental abilities.

§64 Homogeneity and Dimensionality in One-Layer Theories of Abilities: *Independent Cluster Structure (ICS)* and/or *Simple Structure*

Q5 Compare and contrast the two types of "one-layer" theories of abilities, *Independent Cluster Structure* (ICS) and *Simple Structure*.

Q6 What is the definition of a "univocal indicator" and how does it relate to ICS?

Q7 What is meant by the statement that "ICS clusters are conceptualized as ideally *content homogeneous*"?

Q8 How did Varimax and Quartimax help with Thurstone's quest for finding a simple structure?

Q9 In what ways is ICS a special, narrower subcategory of simple structure?

§65 Unidimensionality vs. Multidimensionality on One Layer, an Example

Q10 In what ways can one recognize certain structures of mutual correlations just by visual inspection of the correlation matrix? How can dividing the correlation matrix into submatrices help with this process?

7.2 Confirmatory and Exploratory

§66 The Confirmatory Approach

Q11 Explain the value of the "confirmatory approach" to the investigation of the relational structure of a particular research domain.

Q12 What are the relative advantages of patterned matrices in comparison with path diagrams in the formulation of structural hypotheses for data analysis?

Q13 Near the end of Section §66, in discussing the process of evaluating the goodness of fit of latent variable models, the authors argue that the evaluation should always include a visual check of the residual correlation matrix. Can you see any reasons for this?

§67 Confirmatory vs. Exploratory Simple Structure

Q14 What are the comparative advantages of the confirmatory approach versus the exploratory approach to common factor analysis?

§68 Stronger vs. Weaker Support for Content Homogeneity, by Unidimensionality or by ICS

Q15 What does it mean for an indicator to be univocal? What does it mean for an indicator to be equivocal?

Q16 What is "higher-order factor analysis"?

Q17 How are equivocal indicators related to factorial complexity?

§69 "Hidden Equivocality" in ICS Structures and the So-Called Frame Problem

Q18 What is the frame problem?

Q19 What does it mean for an indicator to be multidimensional?

Q20 What are the implications both positive and negative of an indicator being multidimensional?

Q21 What is "hidden equivocality" in ICS structures?

§70 Dimensionality by the LC Factor Model vs. PCA "Factor Analysis"—Example

Q22 In what ways can factor unidimensionality be useful as a formal model of semantic-content homogeneity?

Q23 Comment on the comparison in Table 7.7 between the one-factor latent common factor model of the first four subtests of CERAD test and the PCA one-factor model of the same data. What makes the LCF model so much more powerful than PCA in identifying the actual correlational structure of the data?

Q24 Which is better, deterministic component/factor scores or factor score indeterminacy?

Q25 How does PCA "explain variance"? What is the difference between the way PCA explains variance and the way the Latent Common Factor Model explains variance?

Q26 What kind of variance do we wish to explain in order to understand correlational structure?

Q27 Why do the authors maintain that communality in PCA is meaningless?

Q28 How many components are needed for PCA to explain all the variance in a covariance matrix or a correlation matrix?

Q29 What is factor score indeterminacy?

Q30 Why is it that factor scores cannot be determined precisely, but only estimated from regression?

7.4.2 B. Exercise Questions

7.4.2.1 Four Analyses Using the Stata sem Procedure

The first several examples in Chapter 7 use the correlation matrix from Table 7.1 for eight psychological variables. Reading this data into Stata for analysis with the Stata sem procedure is accomplished with the following syntax.

```
*** Read the correlation matrix of the
*** eight psychological variables
clear                              // Clear the workspace
ssd init v1 v2 v3 v4 v5 v6 v7 v8 // Variable names
ssd set observations 200           // Number of observations
ssd set correlations /// Correlation matrix
 1 \ ///
.50  1 \ ///
.52 .40  1 \ ///
.43 .54 .50  1 \ ///
.38 .48 .33 .41  1 \ ///
.19 .10 .18 .27 .10  1 \ ///
.20 .10 .20 .24 .10 .89  1 \ ///
.22 .20 .28 .38 .23 .43 .38 1
```

After clearing the workspace, the variable names are given, the number of observations is specified, and the correlation matrix in lower triangular form is brought into Stata. A common initial baseline comparison model is the unidimensional or single-factor model. The sem command uses a factor F with the eight variables as indicators, giving the standardized estimates. The two estat commands are used to obtain the goodness of fit statistics and the correlation residuals. The results of the analysis are given in Table 7.2.

```
*** Eight indicators, one factor,
*** Confirmatory factor analysis, Table 7.2
sem (F -> v1-v8), standardized
estat gof, stats(all)
estat residuals
```

The initial two-factor analysis is based on the initial path diagram (Figure 7.5) and the pattern hypothesis for the corresponding confirmatory two-factor ICS. This analysis can be obtained with the following Stata syntax specification, and the results are given in Table 7.3.

```
*** Eight indicators, two correlated factors,
*** Independent Cluster Structure
*** Confirmatory Factor Analysis, Table 7.3
sem (F1 -> v1-v4) (F2 -> v5-v8), standardized
estat gof, stats(all)
estat residuals
```

The Stata sem command specifies that the factor F1 is measured by indicator variables v1 through v4, and the second factor F2 is measured by indicators v5 through v8. A standardized analysis is requested, as are the goodness of fit and residual statistics. Given the very low factor loading estimate for item v5 on factor F2, the structural hypothesis was modified to permit v5 to load on both F1 and F2. The following Stata syntax was developed to accomplish this. However, Stata's sem program failed to properly converge for this model leading to possibly untrustworthy results. The results from estimating the same model with Mplus are used in Table 7.4 (see the information on Mplus later).

```
*** Eight indicators, two correlated factors,
*** Simple Structure with Cross-loading
*** Confirmatory Factor Analysis, Table 7.4
sem (F1 -> v1-v5) (F2 -> v5-v8) , standardized
estat gof, stats(all)
estat residuals
```

When an indicator loads on multiple factors, or on a different factor than expected, it is common to exclude the item thought to be aberrant from the analysis entirely. Although this tactic raises important concerns that should probably be addressed more carefully, the analysis implementing this approach is given below. Here item v5 is completely removed from the Stata sem command, and the results in Table 7.7 are obtained.

```
*** Seven indicators, two correlated factors,
*** Independent Cluster Structure
*** Confirmatory Factor Analysis, Table 7.6
sem (F1 -> v1-v4) (F2 -> v6-v8) , standardized
estat gof, stats(all)
estat residuals
```

Exploratory Factor Analysis Using the Stata Factormat Procedure

Instead of using confirmatory independent clusters or simple structures, the correlation matrix of interest is sometimes analyzed using so-called EFA. The factormat procedure in Stata can be used to conduct exploratory maximum likelihood common factor analysis using correlation matrix input. This requires a somewhat different process for inputting the correlations as a named Stata matrix which can subsequently be used by factormat. The following Stata syntax can be used for this purpose.

```
*** Read the correlation matrix
clear
matrix C = ///
( 1 , .50, .52, .43, .38, .19, .20, .22 \ ///
.50, 1 , .40, .54, .48, .10, .10, .20 \ ///
.52, .40, 1 , .50, .33, .18, .20, .28 \ ///
.43, .54, .50, 1 , .41, .27, .24, .38 \ ///
.38, .48, .33, .41, 1 , .10, .10, .23 \ ///
.19, .10, .18, .27, .10, 1 , .89, .43 \ ///
.20, .10, .20, .24, .10, .89, 1 , .38 \ ///
.22, .20, .28, .38, .23, .43, .38, 1 )
```

After creating the correlation matrix, the following factormat syntax can be used to conduct a maximum likelihood factor analysis, obtain the correlation residuals, and carry out both Varimax and Oblimin rotations. The factormat command retrieves the correlation matrix C with a sample size of 200, names the variables, extracting two factors with maximum likelihood estimation. The results are given in Table 7.5.

```
*** Two factors exploratory factor analysis Table 7.5
factormat C, n(200) names(v1 v2 v3 v4 v5 v6 v7 v8) ///
factors(2) ml            // Two factors, maximum likelihood
estat residuals          // Correlation residuals
rotate, varimax          // Varimax rotation
rotate, oblique oblimin  // Oblimin rotation
```

Exploratory Factor Analyses (EFA) and Principal Components Analyses (PCA) Using Stata

We can focus on just the first four variables in this dataset to address comparisons of the EFA and PCA solutions for multivariate data. The first step here is to read the smaller 4×4 correlation matrix as a named Stata matrix.

```
*** Four indicators for EFA and PCA
clear
matrix D = ///
(1 , .50, .52, .43 \ ///
.50, 1 , .40, .54 \ ///
.52, .40, 1 , .50 \ ///
.43, .54, .50, 1 )
```

The following syntax uses two factormat commands to estimate single factor common factor analysis or single component PCA, along with the corresponding correlation residuals. The results are presented in Table 7.7.

```
*** One factor ML and PCA with residuals Table 7.7
factormat D, n(200) names(v1 v2 v3 v4) ///
 factors(1) ml                          // ML
estat residuals
factormat D, n(200) names(v1 v2 v3 v4) ///
 factors(1) pcf mineigen(0)            // PCA
estat residuals
```

Notice that with Stata's factormat for PCA, it is also necessary to use the mineigen(0) specification to guarantee that the designated number of components are extracted. The above syntax extracts just a single component, but one can extract two or three components with the following syntax. The results for the two are given in Table 7.8. By combining the results from one LCF and one PCA component (Table 7.7) with the results from two and three PCA components, one has the information needed for Table 7.9.

```
*** Two PCA components Table 7.8
factormat D, n(200) names(v1 v2 v3 v4) ///
 factors(2) pcf mineigen(0)             // Two components
rotate, varimax
estat residuals
*** Three PCA components for Table 7.9
factormat D, n(200) names(v1 v2 v3 v4) ///
 factors(4) pcf mineigen(0)             // Three components
rotate, varimax
estat residuals
```

Mplus Commands for CFA Models for Tables 7.2–7.4 and 7.6

The Mplus input syntax file, Cogdata.inp, to estimate the four CFA models corresponding to those estimated earlier with Stata sem is given here along with the associated data file. Following an optional title, the input file gives the name of the data file, Cogdata.txt, and indicates that it consists of a (lower triangular) correlation matrix. The number of observations (200) is specified, followed by the variable names, and designation of maximum likelihood estimation. Each of the lines in the MODEL: command gives the syntax to estimate one of the CFA models. Only one line is used at a time; unused lines are "commented out" by placing an exclamation point at the beginning of the line. Finally, the OUTPUT: command requests standardized estimates and residuals.

Cogdata.inp:

```
TITLE: Confirmatory Factor Analysis (CFA) Models
DATA: FILE IS cogdata.txt;
 TYPE IS CORR;
 NOBSERVATIONS=200;
VARIABLE: NAMES ARE v1-v8;
ANALYSIS: ESTIMATOR IS ML;
MODEL:
 F BY v1-v8*; F@1;                        ! Table 7.2
! F1 BY v1-v4*; F2 BY v5-v8*; F1-F2@1;  ! Table 7.3
! F1 BY v1-v5*; F2 BY v5-v8*; F1-F2@1;  ! Table 7.4
! F1 BY v1-v4*; F2 BY v6-v8*; F1-F2@1;  ! Table 7.6
OUTPUT: STDYX RESIDUAL;
```

Cogdata.txt:

```
  1
.50  1
.52 .40  1
.43 .54 .50  1
.38 .48 .33 .41  1
.19 .10 .18 .27 .10  1
.20 .10 .20 .24 .10 .89  1
.22 .20 .28 .38 .23 .43 .38 1
```

The results of the Stata and Mplus procedures were generally very close and the Stata results were used in the Chapter 7 tables. However, because of a convergence problem with Stata sem for Table 7.4 model, the results from Mplus were used in this particular chapter table only.

References

Bentler, P. M. (2007). Can scientifically useful hypotheses be tested with correlations? *American Psychologist*, *62*, 772–782.

Beretvas, S. N., & Furlow, C. R. (2006). Evaluation of an approximate method for synthesizing covariance matrices for use in meta-analytic SEM. *Structural Equation Modeling*, *13*, 153–185

Bouchard, C. E., Shephard, R. J., & Stephens, T. E. (1994). Physical activity, fitness, and health: international proceedings and consensus statement. In *International consensus symposium on physical activity, fitness, and health, 2nd May 1992*, Toronto, ON, Canada. Human Kinetics Publishers.

Coan, R. W. (1964). Facts, factors, artifacts: the quest for psychological meaning. *Psychological Review*, *71(2)*, 123–140.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, *105*, 317–327.

Dennett, D. C. (1987). Cognitive wheels: the frame problem of AI. In Z. Pylyshyn (ed.), *The robot's dilemma*. Norwood, NJ, Ablex.

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika*, *2*, 41–54.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200.

McDonald, R. P. (1985). *Factor analysis and related methods*. Lawrence Erlbaum, Hillsdale, NJ.

Neuhaus, J. O. Wrigley, C. (1954). The quartimax method: an analytical approach to orthogonal simple structure. *British Journal of Statistical Psychology*, *7*, 81–91.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201–292.

Štochl, J., Boomsma, A., Tomesova E., & Kovar K. (2006). Structural equation model of motor symptoms of Parkinson's disease. *Studies in Physical Culture and Tourism*, *13*(Supplement) 99–101.

Thurstone, L. L. (1935). *The vectors of mind*. Chicago, University of Chicago Press.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, University of Chicago Press, Psychometric Monographs no. 1.

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago, University of Chicago Press.

Vernon, P. E. (1971). *The structure of human abilities*. London, Methuen.

Wear, C. L. (1951). The evaluation of attitude toward physical education as an activity course. *Research Quarterly. American Association for Health, Physical Education and Recreation*, *22*(1), 114–126.

8

Metrology: Scientific, Applied, and Legal

An Introduction with Implications for Human and Social Sciences

CHAPTER MENU

8.1 Purposes of Metrology, the Science of Measurement, 316
8.2 Brief History of International Metrology and Its Relevance to Human Sciences, 318
  8.2.1 The Ferguson Committee, 320
  8.2.2 Why Is Metrology Missing in the Measurement Literature of the Human Sciences?, 322
8.3 Scientific Metrology, 322
  8.3.1 International Scientific Metrology, 323
  8.3.2 Why Was International Metrology Day, 20 May 2019, so Significant?, 325
  8.3.3 Why Replace Physical Artifacts with Fundamental Constants of Nature?, 325
  8.3.4 Examples: How the First Three SI Base Units Are Defined Scientifically, 325
    8.3.4.1 The Second, 326
    8.3.4.2 The Metre, 326
    8.3.4.3 The Kilogram, 326
8.4 Applied and Legal Metrology, 327
  8.4.1 Applied Metrology Purposes and Key Organizations, 328
  8.4.2 Legal Metrology Purposes and Key Organizations, 328
8.5 A Manifesto for Establishing International Metrology in All of the Sciences, 329
  8.5.1 Extending the VIM for the Human Sciences, 329
  8.5.2 A Sociocognitive Caution About Properties in the Human Sciences, 331
  8.5.3 In Human Sciences, Applied Instrument-Scale Development Comes Before Unit Definition, 331
8.6 Properties: Extensive and Intensive Metadomains, 332
  8.6.1 Extensive, 332
  8.6.2 Intensive, 333
  8.6.3 Critique of the Extensive and Intensive Metadomains, 333
8.7 Ostensive Metadomain, 334
  8.7.1 Ostensive Definition, 334
  8.7.2 Ostensive Performance Demonstration, 335
  8.7.3 Assessing Criterion Performance *Ostensively* at a Capstone Level or a Competition, 335
  8.7.4 Ostensive Demonstrations Entail Reorganized Teaching and Assessment Roles, 338
  8.7.5 Outline for a Possible Ostensive Demonstration Metadomain for Human Sciences, 338
8.8 Summary and Concluding Thoughts, 339
  8.8.1 Lack of International Metrology Citations in Human Sciences Measurement Work, 339
  8.8.2 Looking Ahead: We Will Not Sit Around Waiting, 340
  8.8.3 Open Skills Network and Rich Skills Descriptions, 340
  8.8.4 Putting Rich Skills Sharing "in Context," a Comparison to the History of SI Units, 341
  8.8.5 As a Send-Off, 342
Acknowledgments, 342
8.9 Study Questions, 343
References, 345

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

8.1 Purposes of Metrology, the Science of Measurement

International cooperation in commonly accepted "weights and measures," a result of the growth of international *Metrology*, cuts through and skirts around international competition and international conflicts in a variety of ways. It creates a remarkably cordial and cooperative "safe zone" where the scientists lead us all in agreeing to use a common set of coherently connected metric units that enable scientific results to be compared and new findings to be developed cooperatively. Consider the base Unit for Time in the International System of units, (or SI—the metric system). Its SI unit is the second, and when combined with the base unit, length or distance brings together our schedules and locations throughout an interconnected world. The base units (and derived units made from them) are products of Scientific Metrology. It defines them and requires proof that any instrument measuring them traces to the scientific definition and has an uncertainty low enough to serve its purpose. Applied or Industrial Metrology meets industry's demand for good instruments and proof of their accuracy and fidelity. Industries worldwide that follow scientific standards have put clocks or watches (often incredibly small and powerful ones) in the possession of most everyone; usually, on the wrist or cell phone, and with it, GPS information. This is so in the developed parts of the world and is rapidly changing in the less developed. Legal Metrology has followed regulations, and has lobbied for better ones, to provide laws and tools to prevent and challenge fraudulent or careless measurement instrument developers. This is essential for keeping public trust high in the accepted common units. Trust in accurate measurements is fundamental in order for economies to thrive through trade.

Metrology is the ***fundamental science*** *of measurement*, and it has applied (or industrial) and legal branches as well. They are segregated appropriately from each other but collaborate effectively. A set of international organizations for Scientific, Applied, and Legal aspects of Metrology are introduced in this chapter, along with what they do. References are provided to rich sources of interesting and important documents.

Our purposes in this Metrology chapter are to:

1) Provide a brief review of the history of the metric system including the example of STEM[1] scientists coming together across disciplines to share common units and understanding, and to support international organizations set up to assure scientific, applied, and legal support for continued sharing and collaboration. Also we wish to explore ways to facilitate cross-discipline communication so researchers and practitioners in human science fields beyond medicine, including psychology and education, can access and understand international Metrology, and ultimately become a part of it and obtain its scientific, social, and economic benefits.
2) Connect future psychometricians with the remarkable resources of international Metrology, and how fundamentally these resources can reframe the types of questions currently being pursued without reference to international Metrology.
3) Offer a possible new metadomain of properties that might become measures in addition to the Extensive and Intensive metadomains of physical and chemical properties. This will be

1 This is an acronym for Science, Technology, Engineering, and Mathematics. The acronym "STEM" refers to the historical fact that the dominant players in *Science* have been Physicists, Chemists, Mathematicians, and Engineers. Physical and chemical units comprise the seven base units from which the SI system grew.

approached by introducing Ostensive properties of *human objects*, and how to define and demonstrate them in a way that can stand alongside the Extensive and Intensive properties of STEM domains, leading to the definition of new easy-to-interpret human scales of progressive attainment by providing frequently repeatable measurement on high-quality and user-friendly instruments.

4) Offer a user-friendly and learning-centered framework for developing and using measurement instruments in psychometrics, anthropometrics, and related fields in the human sciences. Growth in the use of measurement is expected to far exceed current visions of how to use testing in a much less compulsory and opaque manner than current testing systems are accustomed to providing, often by government mandate for administrative accountability. It is hoped that path-finding visionaries will arise to contribute to local, regional, and international efforts during their careers, adding important learning-oriented units and scales in their individual branches of human science. Perhaps future developments will hasten the day when national and international Metrology programs will support the scientific development, and applied and legal dissemination, of such instrument scoring and feedback systems—even someday having internationally approved models and scales of progressive attainment, and other applications of measurement in the human sciences.

This book does not use the word "Measurement" in its title, and the original 2010 manuscript for the book does not use the word "Metrology" anywhere. Yet measurement is implicit in the term psychometrics that appears in the book's title. Metrology is the science of measurement. There are numerous, vigorously-growing organizations of cooperative international metrology. Since 1875 at the International Metre convention, these metrology organizations have embraced the domains of physical and chemical sciences, engineering, and much in biology and medicine. Metrology has progressively provided new definitions of quantitative units for properties of objects found in the domains of these developing sciences and technologies. Applied Metrology has progressively provided better, smaller, and less expensive instruments.

Clashes of opinion have occurred both within and across STEM and human science disciplines. Some professionals in STEM disciplines may still prefer to discount or ignore the developments in the measurement of human physical, cognitive, and emotional properties. This discounting occurs even in newer areas such as functional MRI and quantitative cognitive electrophysiology. Unless future Nobel Prize-level work establishes more accessible and scientifically convincing ways to demonstrate game-changing progress — such as valid scales of learning and growth, or medical breakthroughs in healing — the International Committee for Weights and Measures (CIPM)[2], which operates under the International Bureau of Weights and Measures (BIPM)[3], is unlikely to publish human science measurement scales for global standardization within the SI system. As this chapter shows, however, the International Vocabulary of Metrology (VIM 4 and more likely beyond) has the opportunity and willingness to accept modifications applicable to "*all fields of science and technology*," as they prove themselves. Importantly, even now, the current definition of Metrology aspires to serve all fields, which is expected progressively to widen its current reach beyond the STEM disciplines.

2 CIPM is the acronym for the *Comité International des Poids et Mesures,* the French name for the International Council of Weights and Measures. It promotes world-wide uniformity in units of measurement by direct action or by draft resolutions.

3 BIPM is the acronym for the *Bureau International des Poids et Mesures,* the French name for the International Bureau of Weights and Measures which has the responsibility of maintaining international uniformity through the CIPM.

8.2 Brief History of International Metrology and Its Relevance to Human Sciences

International collaboration in weights and measures is a broad and fruitful framework to consider for those in the human sciences. It stands in sharp contrast to the failed opportunity for international collaboration as related in the story of the Ferguson Committee in 1932 England (explained in Chapter 1, Section §6). Both the physicists and the psychologists from different sections of the British Academy of Sciences who served on the Ferguson Committee were well aware of progress in the metric system—internationally accepted measures of base units and derived units. An international treaty in 1875, with the United States as one of 17 original signers, established the internationally owned organization, Bureau International des Poids et Mesures (BIPM—French for "International Bureau of Weights and Measures"). In 1889, the first platinum and iridium standards for a meter, and a kilogram, were approved internationally and copies of the physical artifacts were shared.

But the meter and kilogram had been defined 90 years earlier. Ken Alder (2003) published an excellent and interesting historical research work titled *The Measure of All Things*. In it, he tells the remarkable story of the savants who established the international standard for the meter and later, the rest of the metric system. In the process, they became known as scientists rather than savants, and an international communicating scientific community was established. Even our familiar Least Squares method came out of it as a strong tool for dealing with the management of error, something now carefully specified in international Metrology literature through the Guide to the Expression of Uncertainty in Measurement (GUM; International Bureau of Weights and Measures, & International Organization for Standardization, 2023).

The need for a common system of weights and measures was acute in pre-revolutionary France. Alder explained that the *Ancien Regime* (the 'Old Regime', or the political and social system of France prior to the French Revolution in 1789) listed a staggering 250,000 units of measures and weights, often differing not just between provinces and districts but even between nearby towns. The Enlightenment thinkers saw an opportunity to create an international set of standard measures, opening up international trade, commerce, and scientific exchange. The idea was hatched to make the basis of the "metre" (as spelled then and now in France) one ten-millionth of the distance from the equator to the north pole. Such a measure of the world we all share offers a persuasive reason to agree internationally on a first standard for length/distance, and later measures for weights and other quantities. It was correctly believed that such standardization would galvanize communication and commerce. Thomas Jefferson, as US Secretary of State under George Washington, was in communication as the French sought international collaboration. Jefferson was originally behind defining the metre as the distance a pendulum travels in one second, at the 38th parallel (near his Montecello). The 45th parallel was considered later by the French congress. Jefferson agreed later to the other standards when he was President. But the United States as a whole did not adopt them. But most US scientists did, frequently along with other systems. Today, US manufacturers use it for international products, but only the United States, Liberia, and Myanmar have not converted to it fully.

In these early deliberations, the great French statesman Talleyrand convinced his legislature that 1/4 of the circumference of the earth was the best option before them. The pendulum depended on another unit, the second, and also on the longitude; and could be influenced by the mass of nearby mountains. He also orchestrated passing into law that the various units of the new system would all be linked into a rigorous integrated system and that it would be *decimal-based*, to improve the ease of calculation from systems based on fractions of 12 and 60. The metric system really gained

momentum after the French Revolution, and Jefferson, then president, was sent a prototype kilogram, but it was lost because the boat carrying it and the man carrying it were captured by pirates in the Caribbean and never recovered.

The meridian running through Paris, on up to Dunkerque in the north, and Barcelona in the south, had been measured before, and this portion of the Meridian would be used to estimate the entire 1/4 of the earth's circumference. However, new, much more accurate measures were needed to make this the world's most precise estimate to date. A new instrument "over ten times more precise than the last ones" was selected and two astronomers, Jean Delambre and Pierre Méchain, were each to take a team, north from Paris to Dunkerque or south from Paris to Barcelona, respectively. After each team had returned to Paris, the plan was to combine the completed survey measures and estimate the quarter circumference. They planned to put the combined data into the hands of distinguished mathematicians: Laplace, Legendre, the noted savant Lavoisier, and others. Of interest to psychometricians today, mathematician Carl Friedrich Gauss was at the French Academy of Sciences and worked with the data. Gauss and Adrien-Marie Legendre had a dispute concerning discovery of the least squares method. Either way, it became the foundation of modern statistics.

This progress toward an unprecedented international system of units probably wouldn't have happened at another time than before and during the French Revolution. The Revolutionaries upset traditions and permitted it, but definitely complicated things and endangered even the lives of the savants. Even the great Antoine Lavoisier was beheaded, unpopular for his years of running the hated tax collecting system. But the climate of great change in governance opened minds to the possibility of other big changes affecting the whole world. Despite severe obstacles along the way, both Delambre and Méchain accomplished their work, and the metre was accepted on June 22, 1799 within France. International efforts to bring about international communication and cooperation preceding and following this date were successful, and now the International System of Units is a remarkable story of international cooperation for the common good of all.

From puzzling inconsistencies in a small part of Méchain's work, scientists learned that error was unavoidable in the best scientific work, and how to deal with it. The subtitle of Alder's book is *The Seven-Year Odyssey and Hidden Error That Transformed the World*. There was in 1799 no theory of error nor how to handle it. Fortunately, great mathematicians and scientists, astronomers, and mathematicians were on the Academy of Sciences panel and together masterminded the Metric System. The Least Squares method was applied and became foundational in part because of its use on this highly visible project. Unfortunately, Méchain's very well-measured observations were over a mountainous part of France and Spain that affected his instruments. Also, the earth itself is not a perfect sphere. Méchain thought he made errors because he was off 3.2 degrees of arc in one section. So, he went back over his notes repeatedly trying to overcome the terrible guilt for that "error." It haunted him, and he eventually died on a subsequent mission to extend his Meridian and perhaps replace the erroneous section. After his death, his notes adjusting the numbers were found in his papers by Delambre, who correctly saw that dealing with errors was going to be a continuing part of the scientific endeavor. Today's meter is not exactly one ten-millionth of 1/4 of the circumference of the earth. An error about the size of 0.2 mm, less than three sheets of paper, exists in today's meter. But error and how to handle it became a part of the science of measurement and Méchain's anguish and guilt were unnecessary.

In 1960, all reference to the irregular and inconsistent circumference of the earth was forgotten and the platinum-iridium prototype metre guarded at BIPM was replaced based on the wavelength of the element Krypton 86. In 1983, it was given its current definition; based on the distance light travels in a vacuum in one second.

8.2.1 The Ferguson Committee

The metric system and its use in science were well established by 1932, the year the Ferguson Committee of the British Association for the Advancement of Science began its deliberations. Chapter 1 discussed the seven-year debate between physicists and psychologists, and by implication, between hard scientists and people in many other fields of the human sciences. Over seven years of debate, the physicists would not budge and allow that psychologists could develop real scientific measurements. They would not deviate from N.R. Campbell's (1920, 1928) long-standing definition of scientific measurement. To deny that Stevens' Sone scale of perceived loudness was true measurement required an argument that hinged on the necessity of an operation of concatenation before the behavioral scientist could claim his scale represents true scientific measurement.

In 1939, when the Ferguson Committee rendered its opinion on the Sone Scale, the second had been approved by International Metrology for some decades, and the meter and kilogram copies were being used daily in manufacturing and commerce. All three could be demonstrated by concatenation to reproduce sums and ratios of different quantities of correctly measured and weighed objects. It seemed obvious at that time to the physicists that "true, scientific measurement" required the operation of concatenation. But many chemical and additional physical properties were being developed at that time and were already, or soon would be, approved as SI units—intensive properties—that do not support concatenation. Consider an intensive unit as simple as Density, that is, mass divided by volume. To apply Campbell's concatenation rule would also have voided this new unit as well as it voided Stevens' Sone scale. Density and many other units derived from the base units, and accepted as intensive units by the metrologists, would also not have been accepted. This is an example of where cross-disciplinary collaboration could have occurred but did not. It is not clear that the Sone scale should have been approved at that time, but it should not have failed to be approved for an erroneous scientific reason. Concatenation is surely *sufficient* to demonstrate measurement on a ratio scale, including having additive units, but is it a *necessary* condition for establishing a ratio scale? It is not, as shown by several intensive combination measures, that qualify under strict Metrological conditions as SI units and yet do not pass the test of concatenation.

As related in Chapter 1, Section §6, Stevens' (1946) response did not appear until six years after the Ferguson Committee's final report. When it was published in *Science*, it became one of the most influential measurement papers of the 20th century. Although it is an easily understood clarification of scale properties, as stated by Narens (2002), it "clearly is inadequate for the rigorous founding of a theory of measurement." It is certainly not comparable to the rigor of the axiomatic approach of Krantz et al. (1971) in their three volumes on representational measurement theory.

Norman Cliff is a behavioral scientist who clearly saw the incompleteness of psychology's evidence that its measures are quantitative, not only in Stevens' four measurement scales but also in the much more rigorous Foundations of Measurement of Krantz et al. (1971), as discussed in Chapter 1. His conclusions are found in his article titled "Abstract Measurement Theory and the Revolution That Never Happened" (Cliff, 1992). In it, he argued that the logic of the theory is opaque, inaccessible to most psychologists, and also pointed to the difficulty in providing proofs of theorems of representation and uniqueness for each new psychological scale. He published a book on purely ordinal methods in psychology (Cliff, 1996). In this book, Cliff agrees that the Representational Theory of Measurement has given a pathway, through Additive Conjoint Measurement (ACM), to prove that a particular scale meets the theorems of ACM, and thus is additive.

However, he argues, since many questions psychologists ask are purely ordinal, and the proof of additivity is difficult and often fails, why not use ordinal methods which do not require proof of additivity?

Going deeper on the issue of proving additivity, psychologist and philosopher of measurement, Joel Michell (2008) published a paper titled "Is psychometrics pathological science?" He states: "Pathology of science occurs when the normal processes of scientific investigation break down and a hypothesis is accepted as true within the mainstream of a discipline without a serious attempt being made to test it and without any recognition that this is happening." He argues that this has happened in psychometrics on the issue of whether the scales being used in research are actually quantitative.

Psychometricians have a responsibility to show that their scales are additive before performing the many kinds of statistical operations that assume scales are quantitative in at least the interval sense. Michell in other publications has explained in detail how to do this. It involves checking certain cancellation conditions required by the theory of ACM. However, these cancellation tests are time-consuming and worse still, the conditions are often not met. Human subjects in a sample vary in individual ways that lead to different levels of difficulty in some of the items or tasks on a test—intransitivities—which make a cancellation condition fail. Even a simple ordinal scale must demonstrate transitivity among the responses of all or the great majority of test takers. In regard to Norman Cliff's question that if the research questions are only ordinal questions, isn't it better to use ordinal methods; our answer is that the latent factor statistical machinery and IRT provide so much more inferential power and validational power that there is no need to settle for ordinal methods. The field as a whole has given Norman Cliff similar answers but still has not answered Michell's challenges.

Interestingly, Michell is not the first one to point out that parts of psychology are "pathological." Sigmund Koch (1992) used this term to describe the strange phenomenon of defending the impoverished logic of "operational definitions" of variables by attributing their origin to Nobel Laureate physicist Percy Bridgman. Koch tried for years to get psychologists to see the pathology in this error. In 1927, Bridgman wrote an unfortunate paragraph in his book discussing operational analysis, an approach to analyzing a scientific problem, by looking at the operations used to measure them. Stevens found and interpreted this paragraph wrongly, contributing to the way operationism has been understood for so long in much of psychology. (Petr Blahuš defines himself as rejecting operationism and determinism.) Stevens—the same Stevens who was rebuffed wrongly by the physicists—started almost a century of attributing to Bridgman psychology's favorite shared delusion, that Bridgman originated the idea of the operational definition.

Bridgman wrote and spoke for years against this misuse of his work by psychologists, but he was ignored. He never wrote about operational *definitions*, except to deny that he originated the idea. Sigmund Koch, who edited the scholarly series "Psychology, the Study of a Science" tried afterward to point out the many areas in which psychology fell short. He tried for years to get psychology to accept its own eccentricity regarding the origin of extreme operationalism and other matters, but he was also ignored. Bridgman is still quoted erroneously by books written in the 21st century, many cited in this book, as the originator of operational definitions and strict operationism. Psychology textbooks still teach operational definitions as a reasonable approach to research design. It is not just the physicists who persist in outdated traditions and delusions. Psychologists and others do so as well. This rounds out the failed opportunity for physicists and psychologists to get together. Campbell failed to advance positive ways for scales like the Sone scale to be improved, and Stevens, who must have heard some of Bridgeman's pleas to stop misrepresenting his views of science, continued to promote this brand of operationism.

8.2.2 Why Is Metrology Missing in the Measurement Literature of the Human Sciences?

The simplest explanation is that the scope and value of International Metrology have both been evolving so rapidly that perhaps only a minority of STEM scientists themselves are up on it. So much of it is very new and has not sufficiently been put forth to the professionals in the STEM domains, let alone to the human scientists. So much in modern Metrology dates back only to 1960, and ever so much more in the last 30 years. Before 1960, there were only 3 or 4 base units. The Joint Committee on Guides to Metrology (JCGM), a key component in International Applied Metrology, only dates back to 1997.

In reviewing books on measurement in psychology and related fields, including educational measurement, we found a number of passages where the authors were willing to discuss how physical measurements take place, but the authoritative Metrology references on those units were rarely even mentioned as an aside. A clear understanding of scientific Metrology and its history is quite enlightening as an example of how science progresses. It is reflected in improvements in the measurement units associated with the measured property. These free reference works are readily available from references given in this chapter and elsewhere. An example is the discovery by Josephson (1962) and of what is now called the Josephson effect, which is a phenomenon that occurs when two superconductors are placed in proximity with some barrier or restriction between them. This effect has enabled the charge of a single electron to be estimated, thus updating the electrical units in the SI system, by tracing them back to the charge of the elementary particle.

8.3 Scientific Metrology

We turn next to explaining the nature of International Scientific Metrology, and how it spawns Applied and Legal Metrology. Advancements in science lead to new units, or better definitions of known units. Advances in science also motivate new instrument development. But the application process is complex. Will enough manufacturers gain the know-how and tools to produce enough instruments? Will their metricians be able to evaluate and provide data on the uncertainty of different levels of instruments for various purposes? It is a massive social and economic undertaking. On the legal Metrology front, those employed in the effort must respond to legislation that happens to get passed involving SI measures. Others in legal Metrology may be figuring out and lobbying for the most socially, economically, and metrologically sound legislation. Add to that trying to coordinate the adoption of consistent and fair regulations across nations, then responding to new social and economic challenges like global climate change and health due to pollution. These require the right measurement technology and legislation to get it deployed, to assess progress, or to detect illegal activities. The possibility of conflict of interest is present in many ways, particularly if there is no separation and accountability between the enforcement on the legal side and the big money available to sway regulators.

One recommendation of this book is that Educational Measurement could greatly benefit from not mingling the science, the applied business, and the legal implications as blithely as is now the case. It is important for human scientists to understand the disciplines for each of the three aspects of Metrology as they are emerging from the fields that are hundreds of years farther along in their Metrology enterprises. Getting to this point from where we are now is not going to be a swift process.

8.3.1 International Scientific Metrology

The International Bureau of Weights and Measures (BIPM) was set up by the Metre Convention signed in Paris on 20 May 1875 by seventeen states during the final session of the diplomatic Conference on the Metre. Two metric units were approved at the treaty conference (the metre and the kilogram) and 40 copies of the Prototype Kilogram and Metre were made in Paris and distributed to other countries. The BIPM has its headquarters near Paris, on the grounds placed at its disposal by the French Government. Its upkeep is financed jointly by the Member States of the Metre Convention. The following several paragraphs are from BIPM's website:

> "The task of the BIPM is to ensure worldwide unification of measurements; its objectives are to:
> - represent the worldwide measurement community, aiming to maximize its uptake and impact,
> - be a center for scientific and technical collaboration between Member States, providing capabilities for international measurement comparisons on a shared-cost basis,
> - be the coordinator of the worldwide measurement system, ensuring it gives comparable and internationally accepted measurement results."

The BIPM operates under the exclusive supervision of the International Committee for Weights and Measures (CIPM) which itself comes under the authority of the General Conference on Weights and Measures (CGPM) and reports to it on the work accomplished by the BIPM.

What is Metrology? (This definition, with formatting, is taken from the BIPM website and given a border)

Metrology is the science of measurement,
embracing both experimental and theoretical determinations at
any level of uncertainty
in any field of science and technology

Measurement science is not, however, purely the preserve of scientists. It is something of vital importance to us all. The intricate but invisible network of services, suppliers, and communications upon which we are all dependent relies on Metrology for their efficient and reliable operation. For example:

- The economic success of nations depends upon the ability to manufacture and trade precisely made and tested products and components.
- Satellite navigation systems and international time correlation make accurate location possible—allowing the networking of computer systems around the world, and permitting aircraft to land in poor visibility.
- Human health depends critically on the ability to make accurate diagnosis, and in which reliable measurement is increasingly important.
- Consumers have to trust the amount of petrol delivered by a pump.

Note that this simple and highly general definition of Metrology includes the words "*at any level of uncertainty*." The new SI unit definitions range from uncertainties of 10^{-8} to 10^{-34} for minute nano-quantities to light years for great distances. This progress in the accuracy and scope of measurement has led to the successes of the space age and earth-bound self-driving vehicles, incredible medical devices, and other wonders of the modern technological age we live in. It is also paving the way for new standards offering even greater accuracy—e.g. the new science of nanomaterials

(Jorio and Dresslhaus, 2016). However, the domains of human behavioral constructs neither require nanometer accuracy nor extremely expensive instrumentation within our laboratories, offices, and public venues.[4]

We need instruments possessing levels of accuracy, costs, and utility appropriate to the domains we value most that need good measurement. In the human sciences, we are graduating from paper instruments to online interactive instruments, but we have a long way to go to have international laboratories establish the definition of units, determine appropriate uncertainties for the intended uses, and assure their traceability to international standard definitions. In human sciences, scientists (whether individuals or as groups) get grants and contracts when possible to do research that involves developing instruments. Human sciences are co-mingling Scientific Metrology and Applied Metrology. When selling the use of their instruments, they look very carefully at their nation's and state's legal regulations and liabilities. Thus, what is seen as Legal Metrology internationally is also co-mingled.

Note also that both *experimental and theoretical determinations* are part of Metrology's definition. One without the other cannot supply explanatory power and evidence of both validity and measurement precision. Petr Blahuš uses the term "*substantive*" to refer to the theoretical rationales, and the term "*semantic*" to refer to the meaningfulness and interpretive power inherent in scales with the features of scientific theoretical substance, accuracy, and meaning. He uses *formal* to refer to the mathematical formalizations that underlie our methods. His views are consistent with international principles, and he served on his nation's academy of sciences.

Scientific Metrology has created some of the biggest specialized science laboratories in the world. Member nations all contribute to the work of BIPD, and all nations are encouraged to fund their own national institutes and applied and legal entities. The National Institute for Standards and Technology (NIST) is the US National Metrological Institute. NIST was involved in deep science advancements by collaborating centrally in the recent change of the kilogram, ampere, kelvin, and other SI fundamental unit definitions.

The United States surely has one of the most advanced National Institutes of Metrology in the world. The NIST Lab buildings are themselves a part of its instrumentation. The Watt Lab in Gaithersburg, MD was built many storeys deep into the earth, and rests on bedrock, isolated from the vibrations of adjoining soils. This is needed to exclude extraneous influences and drive error of measurement down as far as possible. The measurement instruments go through their millions of observations in experimental studies to refine standards to deeper levels of accuracy, and every effort is made to remove controllable sources of error. It is useful to examine the scientific instrumentation used at these labs, and the years of refinements and experiments to evaluate and reduce the most miniscule sources of error in the instruments.

An excellent example is the description of the Kibble Balance (formerly known as a watt balance, a metrological instrument that can realize the definition of the kilogram unit of mass based on fundamental constants). The Kibble balance was used at NIST and other-world Metrology labs to obtain the estimate of the Planck constant that replaced the International Prototype Kilogram (IPK) (Steiner et al. 2005). In this publication, a revealing graphic of the instrument shows only one side of the balance, which loops over a wheel to avoid rigid pans and to avoid error from the sweep of the rigid arms of conventional lab balances. Errors from each imperfect and worn-by-use physical weight in the old box of prototype weights are removed by this design. The Kibble Balance uses a single receptacle for sample objects with weights up to 2 kg. The force produced

4 M.S. Dresslhaus had a son Paul Dresselhaus working in the NIST Boulder lab. Paul and five other NIST scientists, including Yi-hua Tang, whose influence led to this chapter, received the DOC Gold Medal Award in 2007 for developing the world's first quantum-based electric standard for ac voltage.

by the weight of each object placed in the cup is balanced by an electromagnet with varying electrical current. This current is precisely measured and converted to joules of force required to balance each experimental weight placed in the balance cup, leading up to the new estimates for kilogram and current. The apparatus was completely re-made part-by-part more than once over more than 10 years to keep reducing sources of error.

8.3.2 Why Was International Metrology Day, 20 May 2019, so Significant?

In November 2018, a set of historic changes were made in the definitions of the seven base units, especially affecting the kilogram, the ampere, the elementary charge of the electron, and the kelvin. These approved changes took effect worldwide on International Metrology Day, 20 May 2019. The 17 original member nations that signed the metre treaty in 1875 have grown currently to 62 members and 58 Associate nations. You are encouraged to bookmark and study the website for the international nongovernmental organization, the *Bureau International des Poids et Mesures* (1998). The BIPD site provides information on all of the SI Units, (SI → Système International), on all the governing organizations of BIPM connected with Applied and Legal Metrology, and associated international organizations associated with standards for various domains. We give you links at the end of this chapter to the websites of BIPM and also, in the United States, to the National Institute of Standards and Technology (NIST). In a refreshingly transparent manner, these cooperating organizations make information freely available to citizens of member nations to learn the importance of Metrology and about the international standards for Scientific, Applied, and Legal Metrology.

8.3.3 Why Replace Physical Artifacts with Fundamental Constants of Nature?

The answers to these questions are found in Stock et al. (2019). This BIPM report is downloadable. In Section 3 of this report (p. 2), it states that there were data showing that using the different physical copies of the kilogram introduced errors. Even the original International Prototype Kilogram (*IPK*) was suffering an accumulation of contamination and wear on the rarely used platinum/iridium object. The history is found in Appendix 3 of the SI Brochure, Version 9. Trying to prevent any form of degradation in the physical IPK in Paris was problematic. Even under highly controlled conditions (actually, only taking it out about every 50 years), it was degrading. The 40 original copies also varied in exact mass, and they, being used more frequently in calibrations in different nations, were found to vary even more. The Metrological *traceability* back to the original standard was becoming less and less accurate. Another problem is that when much lower masses were to be compared and calibrated (centigrams, grams, micrograms, and now with nanomaterials, nanograms), the comparisons had to be linked to the standard through a chain of subdivision, each adding additional uncertainty.

It was determined to replace the IPK with a definition of the kilogram based on an equation using the Planck constant as discussed below. Thus, the *definition* of this critically important base unit was stripped of any single operational method for realizing it. Also, Applied Metrology documents were prepared so that each country could produce instruments using the best and newest technologies as they are developed and calibrated.

8.3.4 Examples: How the First Three SI Base Units Are Defined Scientifically

The next set of definitions follows closely from the latest (9th edition) BIPM "brochure" of the SI units. We show the definitions for three of the seven base international units: *second*, *metre*, and *kilogram*. The definition is quoted in bold for each of the three units; and, also in bold, its effect for two of the three units is given afterward. Note that the *second (s)* definition appears in the definition

of both length and mass, and the definition of length in that of mass. Both the metre and the kilogram use the second in their definitions, and the definition of length is used to define the derived unit, area in meters squared that standardizes the force of the mass being measured.

8.3.4.1 The Second

The second, symbol **s**, is the SI unit of time. It is defined by taking the fixed numerical value of the cesium frequency, $\Delta\nu_{Cs}$, the unperturbed ground-state hyperfine transition frequency of the cesium 133 atom, to be 9 192 631 770 when expressed in the unit Hz, which is equal to s^{-1}.

> *Note: The metre, then the kilogram were the first two base units approved in 1799. Napolean scoffed at the metric system and delayed its implementation, but his successor restored it. In 1832, Gauss advocated for the second as a unit and it was later approved as the 3rd base unit. Because so many other units depend on it in 21st-century defining equations, including the dependence on the second by the original two base units*, the *metre and the kilogram, it is usually listed first in these days of scientific definitions using equations and constants of nature. Note: In these new definitions, for long numbers above 1000, the international BIPM notation uses a space and not a comma "," as the separator.*

8.3.4.2 The Metre

The metre, symbol **m**, is the SI unit of length. It is defined by taking the fixed numerical value of the speed of light in a vacuum, c, to be 299 792 458 when expressed in the unit m s^{-1}, where the second is defined in terms of the cesium frequency $\Delta\nu_{Cs}$. This definition implies the exact relation $c = 299\ 792\ 458$ m s^{-1}. Inverting this relation gives an exact expression for the metre in terms of the defining constants c and $\Delta\nu$.

The effect of this definition is that one metre is the length of the path traveled by light in a vacuum during a time interval with a duration of 1/299 792 458 of a second.

8.3.4.3 The Kilogram

The kilogram, symbol **kg**, is the SI unit of mass. It is defined by taking the fixed numerical value of the Planck constant, h, to be $6.626\ 070\ 15 \times 10^{-34}$ when expressed in the unit J s, which is equal to kg m^2 s^{-1}, where the metre and the second are defined in terms of c and $\Delta\nu_{Cs}$.

The effect of this definition implies the exact relation $h = 6.626\ 070\ 15 \times 10^{-34}$ kg m^2 s^{-1}. Inverting this relation gives an exact expression for the kilogram in terms of the three defining constants, h, c, and $\Delta\nu_{Cs}$.

Note that with this definition, one can solve for a single gram, or any fractional or multiple amounts of 1 kg. No subdivision of a physical artifact is required for comparison. With newly applied Metrology tools, such as industrial versions of the Kibble balance discussed above, for smaller and larger weighings, industries can set any desired weight in normal ranges and the new industrial balance will confirm a sample product's specified weight. Applied Metrologists are working on applied devices for industry and commerce for the burgeoning new field of nanotechnology, and each of them can use the definitions of size and weight and other SI units, using the metrological science definitions. Note also that while necessary for derived measurement in physics and chemistry, this degree of resolution is vastly more precise than is needed at the scale of human accomplishments, traits, and medical conditions.

It is easy and quick to get to the BIPM website and explore it. Look for the SI units, the seven base units, and other detailed information. Download your own pdf copy of the SI manual, 9th edition, of the BIPM if you wish. You may also browse to the NIST in the United States, and download your

own wallet card for the units, read more about the Kibble balance and the new estimate of the Planck constant. For the more serious, download your own copy of the VIM3 (De Bièvre, 2012), International Vocabulary of Metrology—the dictionary most needed by metrologists whose home is their own professional discipline. VIM4 has not been published yet and cannot be quoted, but a draft pdf copy is available for download for interested professionals—or anyone to review.

8.4 Applied and Legal Metrology

The bullet points in the last boxed BIPM paragraph give the quick and easily understandable reasons why there is an Applied Metrology—to get quality instruments manufactured and distributed. Quality means traceability and accuracy to the scientific definition of each unit. Industrial Metrology that applies the scientific units and standards must operate quietly, behind the scenes.

The subtleties of the entire Metrology enterprise with its scientific, applied, and legal infrastructure are understood and explained with clear simplicity by British Metrologist Richard J.C. Brown (2021) who works at the National Physical Laboratory in Teddington, United Kingdom.

> Metrology remains a uniquely important endeavor. A sign of its success and robustness as an intratechnology is that it usually goes unnoticed. This means that it is in danger of being undervalued and underappreciated.

He published this in August 2020, looking forward to Metrology providing a sturdy quality infrastructure to address global challenges of pandemic recovery. Speaking of how we all rely on trustworthy measurements in almost all of daily life, he adds:

> Metrology is the structure that ensures these measurements are stable, comparable and accurate, providing confidence in measurement at a stated level (usually by quoting a measurement uncertainty). When these qualities are associated with a measurement, it reduces waste, allows trade, enables infrastructure to function, technology to advance, the economy to prosper, encourages global agreement, collaboration and trade, and ensures our ongoing health and safety and quality of life.

The basic idea of Legal Metrology is also not difficult. *Legal Metrology* deals with the regulatory requirements of well-established measurements and measuring instruments for the promotion of fair trade and the protection of consumers. To fully understand what has evolved to accomplish these valuable outcomes, especially in this century, would be too extensive for this chapter, but what is sought here is to provide an overview of the interplay of scientific, applied, and legal Metrology. Also, to provide some sources for further information about what has been involved in getting scientific Metrology set up in each country, and what is involved in Applied or Industrial Metrology. Certifying that the scientific standard is actually being measured requires licensing of many different applied instrument developers. A further task is certifying calibration operations to make sure the instruments are accurate and have an uncertainty appropriate to their purpose.

The BIPM must look ahead many years, and it was doing so in 1995. Its scientific officers over different domains foresaw great growth. This growth was expected both in the size of its portfolio of SI units and in the number of nations that would be using instruments to measure to international standards. It also hoped that at least some of the progress foreseen that took effect in 2019, Metrology Day, would be achieved, and it was.

In October 1995, the CGPM commissioned a study. This study was to look ahead and predict the international and national needs for a wider scope of Metrology services. This report *National and international needs relating to Metrology: International collaborations and the role of the BIPM* was published in 1998 and sent to all member nations of the Treaty on the Metre. It is a good unified source to obtain information about the Applied and Legal Metrology functions and organizations that were needed looking forward. This reference for shorthand purposes will be called *Metrology Needs* (1998). Using this document, an overview of major organizations that support the functions of Applied and Legal Metrology are summarized in Sections 8.4.1 and 8.4.2.

8.4.1 Applied Metrology Purposes and Key Organizations

Applied purposes of National Metrological Institutes like the NIST in the United States include licensing manufacturers to develop instruments, and ways to make widely available testing procedures to confirm calibrations. Standards must be promulgated and checked for conformance. The two then-existing organizations were listed as global organizations.

- *International Laboratory of Accreditation Cooperation (ILAC).*
- *International Organization for Standardization (ISO).*

Other regional organizations in Metrology were listed for the Asia/Pacific, European, and North American regions.

Not mentioned in the report was the *JCGM*. This organization was formed in 1997 and took over and expanded some *ISO* functions to better meet unforeseen needs. JCGM deserves mention here, as it now prepares and publishes the current GUM and VIM documents. GUM is *Guide to the Expression of Uncertainty in Measurement*. VIM is the *International Vocabulary in Metrology*. JCGM partners with a number of international organizations, including general, legal, and ISO, and also with discipline-specific groups in electrochemical science, chemistry, laboratory medicine, and physics.

8.4.2 Legal Metrology Purposes and Key Organizations

The primary organization is the International Organization for Legal Metrology (OIML).

From *Metrology Needs* (1998, p. 91), we read *"Governments have over the years enacted a considerable number of laws and regulations that require reliable measurements for their fair and effective implementation. The term 'legal Metrology' is used to cover this field of interaction between regulations and measurement. Legal Metrology originated from the need to ensure fair trade and provide consumer protection, and it continues to place its greatest emphasis on this area. In recent decades, however, the need to protect society in other areas such as health, safety and the environment has led to new laws and regulations, and a broadening of legal Metrology activities."*

It is also the case that many regional and international trade agreements are signed by treaty, and this binds the nations to agreements that demand measurements, and these have consequences; *"... trade agreements between nations and regions now specifically require all signatories to accept the results of measurements and tests performed by any other party."* Despite the fact that not all countries abide by the agreement to make sure the measures are the real measures and are accurate (they have economic interests in not doing so), other countries have to agree to accept their measures. Thus, there is a great need to assure each country that the other country's measurements meet international standards and are handled in fairness.

8.5 A Manifesto for Establishing International Metrology in All of the Sciences

The new book referred to in the title of this Section §8.5 is a collaboration organized by an Italian Engineer-Metrologist, Luka Mari, who served on the panels that produced VIM 3 in 2012. His colleagues, Mark Wilson and Andrew Maul, are psychometricians and educational measurement experts. They enjoyed a 12-year collaboration producing a number of foundational articles that provided background for key chapters in their book, first published in 2021. This excellent book came to the attention of your current authors in May of 2022, as a very welcome contribution. The contents of this new book we refer to as a manifesto are not unfamiliar to the readers of this book. You have seen its impact on validity in Chapter 3.1. The new definition of measurement was broken out into its parts and explained there. It challenges the existing definitions and models of validity and gives credence to the claim your current authors endorse, that in psychometrics the concepts of validity are unsettled conceptually and scientifically. Later in Chapter 3, Petr outlines, in conjunction with Table 3.3 of Subsection 3.4.5.1 of Chapter 3, what kinds of processes are required (at various levels of scientific and mathematical sophistication) to enable the explicit transformation from properties of objects in the empirical world into numerical values that can be interpreted and used in the "Information world" (human users). That term would likely have resonated with Petr's understanding of the logic of the creation of formalized, that is mathematized, variables.

In Subsection 3.4.5.2, he explains how the new understanding of measurement provides a means to convert from an empirical property to a value measuring that property on an instrument scale, for use in the interpretation of a human situation. The replacement with Mari et al.'s definition of direct measurement is most satisfying for Petr's goals. It enables the conversion from an empirical property to a value measuring that property on an instrument scale, for use in interpreting a human situation. The replacement with Mari et al.'s definition of direct measurement is particularly satisfying for Petr's goals. This new Metrology-framed book[5] also assists Chapters 4, 5 and 6 as we get into IRT and the latent factor model, and centrally in Chapter 6, Figure 6.3 where it fits Petr's conceptual models perfectly and strengthens his justifications for the latent factor model.

The term "manifesto" seems to the authors of this book to apply to the Mari, Wilson, and Maul book in its entirety. It comes upon the world of measurement with a demonstration that there is indeed a commonality in the definition and deeper meaning of what measurement is. This shows the way that it can apply across all sciences, not just the STEM sciences. The term "Manifesto" is first used by Mari, Wilson, and Maul (2023)[6]. This is the second edition of their book, and is their final statement, offered modestly.

We revised several of the chapters of this book after digesting the importance of this new definition of measurement. It gave new clarity to the important concepts of this psychometrics book. We are most hopeful that the goal expressed in the title of this book, *Measurement Across the Sciences*, will be strongly hastened by their work.

8.5.1 Extending the VIM for the Human Sciences

The new definition of measurement given by Mari et al. (2021) is explained in more detail starting in Chapter 3.0. The definition is:

5 Before the final editing of this book, we were pleased to learn that the Mari, et al. (2021) book, *Measurement across the sciences: Developing a shared concept system for Measurement*, has now been released in a second edition, Mari, et al. (2023). The second edition is an Open Access e-edition.

6 The term "manifesto" appears only in the second edition of the book, in *Section 8.3.5, Consequences for the Theory and the Practice of Measurement, in Box 8.1, Toward a manifesto for a widespread Metrological Culture*. The second edition is open access.

> *Measurement is an empirical and informational process designed on purpose, whose input is a property of an object that produces justified information in the form of values of that property*. (This is a generalization of the VIM definition of measurement in 2.1 of VIM 3)

In the 2023 eEdition the crucial word "explicitly" is added before 'justified'. This definition and its grounding in proposed modification to the VIM reverberates and challenges many. First among those they challenge is the international Metrology establishment itself. They challenge it to open up its VIM to all sciences and all measurable properties. They point out that the VIM has definitions and assumptions that permit certain sciences, engineering, and mathematics to participate in the VIM. A suffocating assumption is that the only acceptable properties that can be turned into measures are already "quantities." How can international Metrology be *for all sciences and technologies*, *at any level of uncertainty*, when its VIM excludes the work of metricians of various names in human science fields?

They also challenge the measurement literature of educational measurement and psychology. They target it by including Reading Comprehension Ability as one of their three example properties that can be turned into values. In this book, we call that an Ostensive property and use this as a lead-in to placing this metadomain alongside Extensive and Intensive Metadomains. The other key examples Mari et al. carry through their book are length, an Extensive property, and Temperature, an Intensive property.

One can point to examples and nonexamples of reading comprehension ability at different levels of development grades, ages, and subjects. They use length (an Extensive property) and temperature (an Intensive property) as their other two key examples embraced by their new definition. Early measurement icons Campbell and Holder are shown to have implicit assumptions challenged and better ones defined. Representational Theorists who have left out the importance of the instrument and the empirical environment of measuring activities are credited, then upstepped with a broader picture. The new definition shows that real properties interact with the instrument empirically and the instrument produces values. Such empirical connection was needed to bridge from a qualitative empirical world to numerical information—values—in an information world. There is also a need for interpretive information specific to how the measurements are applied in real-world contexts with legal implications. Mathematical proofs are not enough. Joel Michel's approach to representation is shown to miss the point about properties and values. Stanley Smith Stevens is trimmed back from what Petr calls "measurement by fiat", that measurement is application of numbers according to any rule you can think up. But the Nominal, Ordinal, Interval, and Ratio scale types are supported, with much more discipline than Stevens offered, because in Chapter 6, *they establish that the value of a property can be a number on a ratio scale, a number on an interval scale, a number or position on an ordinal scale, or a class identifier like the value A on the nominal scale of blood types A, O, and B*. International Metrology is already open to nominal and ordinal scales as exceptional cases, but overwhelmingly the VIM uses the term *quantities* instead of *values* in identifying a location on a scale.

As to their offering of a way to broaden to all the sciences, Mari, Wilson, and Maul provide an appendix (pp. 263–273), *Appendix A*, *A basic concept system of measurement*. The 63 concepts behind this new definition involve terms that are currently in VIM—or analogous, but many are modified to apply across all sciences. About 1/3 are new concept definitions not currently in VIM.

The new definition can be applied to both direct and indirect forms of measurement to embrace sciences not currently considering involvement in the international system. The very meaning of direct measurement has changed, and in so doing, what Petr was calling indirect measurement is not what they define that term to mean. Petr's description of the evolution of excellent quality measurement in Chapter 3, Table 3.3 is now direct measurement, not indirect. It can be applied to properties that support any level of scale type discussed in Chapter 1. When there is an instrument that does all the empirical and informational processes required

by the definition, the measurement is direct. *Indirect measurement* applies where two or more properties are involved, and a calculation is utilized to find the value, as in Mass/Volume = Density.

Certainly, one of the key issues that will take time to resolve in new editions of VIM is that the primitive term objects and properties in the VIM may not be a good fit in the human sciences of psychology and educational measurement. The terms *objects* and *properties* are both primitives—undefined. In the human sciences, we can go along with objects as referring to people, but *properties* are not going to be as uniform from person to person as any of the SI units are in the objects that contain them. This leads us to a discussion of properties, but first, a caution.

8.5.2 A Sociocognitive Caution About Properties in the Human Sciences

In his masterful work on sociocognitive foundations of educational measurement, Mislevy (2018, p. 2) introduces instead of properties the term *"LCS patterns"* across people with different backgrounds and situations. It is an acronym for linguistic, cultural, and substantive patterns. These are the patterns that may be modelled by the sociocognitively framed version of latent variable modelling explained in his book. Inside an individual's actual mental processing at the time the assessment was made, a given LCS pattern may not a be closely approximated a *priori* rather, within the persons are resources they assemble at the time they engaged with the testing instrument. But we ask this—what if the measurement were repeated, a part of practice with model-based feedback in between two or more assessment events? Would not that lead to evidence that something that could be called a property? Words like "skill" may be used to refer to a developed or developing property of an object of the class "human person."

Mislevy's caution, and this question about the impact of multiple measurements as skills increase, was mentioned in Chapter 5.

One other caution will be mentioned, the fact that SI units are defined separately from any particular sample or any particular application situation. But in the human sciences, there is not yet and maybe will never be a unit that is the same up and down a scale. Instead, in scales of human attainment or difference, the nature of the "unit" at a low compared to a high level of an expertise, wellness, or sickness scale may have widely different interpretations. These differences may be local to the specific application situation. Extensive and intensive units have the same scientific definition at every level of their range, and for such an attribute as length, its ranges extend from just above the quantum level to vast interstellar distances.

8.5.3 In Human Sciences, Applied Instrument-Scale Development Comes Before Unit Definition

Consider the domain of reading. One construct scale could be called Reading Comprehension to refer to a construct that is found in thousands of possible applications across different ages, languages, and at the upper end, domains of professional excellence possessing their own special vocabularies and concepts. The development of scales has no practical way to take in a wide range of ages, vocabularies, and levels of skill at reading texts in diverse and highly professional fields. It is a massive job to link age scales to one another forming a learning progression. So, there is no single scientific definition of a single unit seen in the near future. The progress starts in an applied setting of some level of school or higher education. How will it be possible to find a common scientific definition in any sense comparable to the definitions of SI units, profoundly linked to fundamental constants of nature?

That does not mean that good science cannot be done. Also, the history of SI units in this chapter alone shows that weights and measures in 18th-century Europe were not unified nor defined scientifically until the first two, the kilogram and meter, were defined.

If you look at temperature, the applied situation makes all the difference in the span of values of temperature an instrument spans, and its attention to criterion levels. Consider thermometers for human body temperature. The temperature range of living bodies is pretty small, and different disease or stress situations requires attention to criterion values. Same is true in cooking thermometers. Same is true in metallurgy, where the melting point is a crucial criterion level.

It was 350 years ago that all the physical units were confusing the European world. For human sciences, our fields are not that old. What will our Metrology be like in 100 or 150 years? The older sciences have had time to really get their act together internationally.

It is unwise to claim that Psychology and Education can never have scientifically respectable unit definitions. It is reasonable to suppose that they will achieve them, but they will also be applied, that is, situation-specific, and require different designs of instruments and differently detailed versions of interpretive stories. We now turn to the metadomains behind the three properties Mari et al. chose to illustrate measurement across the sciences, Length, Temperature, and Reading Comprehension.

8.6 Properties: Extensive and Intensive Metadomains

The British metrologist Richard Brown described metrology as being at the level of meta-thought above measurement. He proposed thinking of metrology as "measuring measurement." This emphasizes how the characteristic meta-thought associated with metrology distinguishes it from routine measurement.

In an analogous way, we could name extensive and intensive collections of properties as "*metadomains*." We use the term *metadomain* because *extensive, intensive*, and *ostensive* span across numerous professional domains where people can learn about the work in a domain, its terms, concepts, and skills. Domain analysis and model synthesis is where the development of assessment instruments begins. Objects found in those domains have properties. And as the VIM and Mari et al. (2021) describe it, the properties of objects are what are turned into measurement units through instruments by an intentional design process.

Most scholars trace the terms *extensive* and *intensive* back to Richard C. Tolman, an American Physical Chemist who described them in 1917. However, Wikipedia names the German writer Georg Helm in 1898. These broad metadomain terms have become widely used and taught in science programs in secondary school and college.

Extensive and intensive properties are physical and chemical properties of matter. The terms extensive and intensive are not mentioned in VIM3 or in the SI Brochure, but the International Union of Pure and Applied Chemistry defines them (IUPAC, 2007). We will use their definitions in that which follows.

8.6.1 Extensive

> *A quantity that is additive for independent, noninteracting subsystems is called extensive; examples are mass m, volume V, Gibbs energy G. (IUPAC,* 2007*)*

Consider the first three base units in the SI system, time, length, and mass. All three meet the additivity rule to identify a property as extensive. The rule is to take two identical samples of a particular size and concatenate them. If the result is 2x either of the identical samples, the property is extensive. The SI unit *second* fits this concatenation rule and fits the IUPAC definition, as do the

meter and kilogram, although no instance of the *second* being classed as an extensive unit was found by these authors.

Physics and Chemistry have relatively few extensive properties compared to intensive ones. Other examples of extensive properties and their variables are entropy, S; heat capacity, C; Helmholtz energy, A; and spring stiffness, K.

Some writers note that the amount, size, or weight of an object is not useful in identifying or classifying the property under observation. Most intensive properties carry with them ways of identifying or classifying the substance involved, as in chemical reactions. This criticism seems to apply mostly to length, area, volume, and mass because other extensive properties like Gibbs and Helmholtz energy, and spring stiffness, identify their own domain.

8.6.2 Intensive

> *A quantity that is independent of the extent of the system is called intensive; examples are temperature T, pressure, and chemical potential U. (IUPAC, 2007)*

Note that the Pure and Applied Physical Chemists' manual jumps over property to quantity as though the quality of being quantitative is inherent in all of their identified variables, which except for rare exceptions like hardness, they are.

An intensive property does not depend on the extent of the entity being observed and measured. Intensive properties are useful in telling us something unique about the property being observed. Temperature is a base unit usually mentioned first among examples of intensive properties. Truly, it doesn't depend on the extent of the thing whose temperature is to be measured. Color, tensile strength, density, smell, conductivity—there are ever so many intensive properties of objects to be measured, and they indeed do not depend on the extent of the thing being measured.

8.6.3 Critique of the Extensive and Intensive Metadomains

The careful definitions of the IUPSC cited above do not cause confusion like definitions observed in searching educational materials for definitions and usage of these two metadomain terms. However, the IUPAC definition of intensive is negative; that is, by saying "not extensive," they leave it wide open for units that may have something more than physical or chemical properties. For example, ***toxicity*** is listed as an intensive unit in many lists. But it demands for its definition something extra-physical or chemical—life. Toxicity is defined by the effect it has on living cells, organs, or organisms. You need to observe these effects to measure toxicity. The argument we would make is that a different category of nonphysical properties is needed for properties like toxicity. It may be useful to construct a metadomain for properties that are only manifest in living organisms. As for the VIM concept of "object": the toxicity affects living cells, organs, or entire organisms. Studies of toxicity involve living organisms that use toxicity in subduing prey. Some creatures have developed their own defenses against toxicity, as in the case of clownfish.

We are introducing a new metadomain, ostensive, and it may prove to be a better one for life-science properties in general that are only definable when the objects are living. In this book, however, our goals are more modest. We are only attempting to establish ostensive as a metadomain in the human sciences, and our examples are clustered in the niche of human learning and human growth. This limitation is consistent with purposes 3 and 4 given at the beginning of this chapter. (3 *To offer a possible new metadomain of properties that might become measures*, and 4 *To offer a user-friendly and learning-centered framework for developing and using measurement instruments*).

Every science, whether "hard" or "soft," life, social, behavioral, or human has investigators looking for interesting properties. These investigators observe, classify, name, perform experiments, and collect data on phenomena in domains they spend their lives investigating. Intensive physical or chemical properties are constructs—inherently a part of the objects and actions in the domain being investigated. While living objects are constantly measured and weighed and the blood and fluids and other lab tests are taken, the available SI units are based on physical or chemical properties. When sports or other events are judged by time, distance or height, strength, etc., these measures are physical, but the theories and models to explain them are psycho-physical. We use the term Theory (or model) of progressive attainment to explain in a language that learners and coaches can use to promote progress toward defined criterion levels of performance. This model is made up of terms referring to the physical motions of human bodies and the cognitive processing or other psychological activities of the human mind.

8.7 Ostensive Metadomain

We now offer an initial definition of this proposed metadomain for human-science properties. They are defined using Ostensive Definition and measured using Ostensive Demonstration.

Our senior author introduced us to his fascination with his partially developed concepts of Ostensive Definition and Ostensive Demonstration by often commenting on different forms of the term *Ostensive*. This emphasis on Ostensivity caused us great pondering. We coupled this with pondering the lessons that can be learned from the kinesthetic and sports science domains he investigated as a kinanthropometrist. We began to see the value of Ostensive Definition toward developing true criterion measures at important milestones on a scale. We began to see how Ostensive Demonstration could lead to more motivating, more learner-centered, more coachable scales of learning progression. We saw its use for performance demonstrations like reading out loud or creating a written or spoken or artistically rendered exhibit in school subjects. We hope he would have been pleased with the strong relevance we have found in the Ostensive metadomain for at least a wide swath of useful domains in the human sciences. Ostensive Definition and Ostensive Demonstration have been mentioned in the Preface, and in Chapters 4 and 5. Here we propose a more complete definition and explanation of each.

8.7.1 Ostensive Definition

Beyond d*efinition by pointing to*, Ostensive Definition is defining a criterion by pointing to a set of people who possess it, by the performance they can repeatedly demonstrate, and often under certain environmental circumstances. For example, we may be trying to measure world-class 3-point basketball shooting under strong defense. We have videos. We can point to the pro-basketball stars by name to define the top criterion. We know the task—rules for a legitimate 3-point shot, and they are enforced by the referees. We know defenses vary from game to game so we can observe some number of occasions. This definition only applies to human persons. An apparatus that could rise above the defense and shoot with great accuracy does not have this property. The property may differ in how each body is used to accomplish it between a very tall player compared to a shorter player.

This definition and this example respect Mislevy's cautions about not assuming a proficiency or an ability to pre-exist across persons. We cannot be sure that the property is identical from person to person, but it does not matter. It is not just something in the person that defines the property, it is the success at the task under the conditions of performing it. That is the property being transduced

from the empirical world to a value in the information world. The other way around the sociocognitive uncertainty that the property as defined is not the same from person to person and situation to situation is to let it be well-practiced as in the example.

The property exists by its results where it matters, no matter the exact sequence of movements nor what the pattern of well-practiced skills may be. It is not just its person model, not just its task model, it is the measurement itself. A measure is not interpreted only by reference to the person model. It connects the triad of person, task, and situation and enables us to interpret the value of being above this high criterion.

Ostensive definition is about setting criteria. From the first, educational measurement practitioners have treasured the idea of a decisive criterion measure, a gold standard. Given such a criterion, the correlation of a new instrument with it becomes an index of validity. As discussed in Chapter 3, such external evidence of validity is still desirable. A correlation is not sufficient for validation, and anyway, the "gold standard" criterion score has been hard to find. Note that demonstrating the criterion level through an individual-rated performance is not a correlation, but a witnessed demonstration. The correlation over many demonstrations offers research insights and evidence of external validation. The slowness of psychologists and educational psychologists to embrace design, except in "experimental design" may explain the lack of literature on how to identify a true criterion, and then how to design and develop a criterion measure. We do find advice in the literature that when designing a new instrument, it is wise to try to make it as close to the criterion as possible. The goal of Ostensive Definition is to describe both the persons' capability and the performance accomplishment they can demonstrate as an instance of the criterion performance.

Beyond Ostensive Definition is the task instrument and procedures for using it in assessing levels up to and ideally somewhere beyond the high criterion level. The performance demonstration must be well specified and standardized so that it may be used at other times and places and that the results be comparable.

8.7.2 Ostensive Performance Demonstration

An Ostensive Demonstration is generally of one of two types. First, a performance, is given before an audience of at least one person, (the judges), who evaluates the performance. Ideally, more than one judge and other audience members will observe the performance. Second, it is frequently optimal also to create a linguistic demonstration of knowledge of terms and concepts, as described in (1b) of section 8.7.3 below. This suite of both (1a) a Performance Demonstration and (1b) a Linguistic Demonstration (may be objective test items, not a presentation) protects against the conflation of the two in a test blueprint. This error was mentioned in Chapter 5.1, Validating the Content Aspect, and in the chapter summary.

8.7.3 Assessing Criterion Performance *Ostensively* at a Capstone Level or a Competition

1) Assessing a person's current value (or score) in relation to a criterion is by a judged performance, and not a test with multiple short items. It is assumed that rubrics are prepared for judges to rate the performance. During practice the coaches also give feedback. The constructs underlying each rubric make up the explanatory content of the models of progressive attainment culminating in the criterion, while the details of the rubric indicate levels of difficulty and help in defining the criterion performance. The judged performance may be a capstone assessment (such as a post-baccalaureate degree presentation), but more often, it is intermediate competencies of progressive attainment along the way.

a) Performance Demonstration. The construct levels in the model of progressive attainment can be practiced at levels lower than a degree or certificate criterion using various forms of practice, including objective items. But in the 1a category, they mostly involve performance practice and "coaching"—even in academic areas where ostensive demonstrations like drawing models in specified scientific domains are repeated, critiqued, and revised. Other examples are written products synthesizing knowledge of reasoning competencies with concepts important in specific domains.
b) Linguistic Content. As suggested in Chapter 5.1, this type of Ostensive Demonstration can assess level of mastery using objectively scored linguistic content as preliminary or along with the personal performance. The Ostensive Demonstration of Linguistic Knowledge might apply to an entire domain. The terms and concepts of the domain or subdomain being assessed with multiple items could be made accessible on a map revealing the structure of the domain, and the learner can time and again keep trying to exceed some criterion score for domain language mastery. It does not have an audience but is proctored, at least on the occasion for a criterion performance rather than a practice run. There may be remote proctoring (to avoid cheating). This can be a scaffolding attainment before meeting a criterion, but the most important criterion to attain is assessed in the ostensive performance demonstration as described in step 1a.

2) It is a performance that can be repeated a number of times (as seen later, this will require learner-centered institutions designed for it), with the assumption for 1a that learning and coaching go on between times and that the learner can improve as rapidly as is personally possible. Thus, it can be used to assess progress and coach for next steps. The assumption for step 1b is the same, but the feedback may be automated rather than personal.
3) The performance is *Ostential*. This is a new term created here for the purpose of describing a situation where a person's presentation reaches the level of the criterion, and receives approval from interested observers. This outcome is sought, as it is motivating to keep the presenter working for even higher levels on subsequent performances. As in step 5, the model of progressive attainment always provides motivational feedback to encourage the learner to *improve and try again*.
4) The performance is in miniature a component of possible careers that are open to only the ones who reach and exceed the criterion. Thus, it can become a personal goal and not an unpleasant duty. The domain analysis that led to the assessment provides pathways to much higher goals that incorporate lower and intermediate milestones.
5) The instrument that administers the Performance Demonstration multiple times has a model of progressive attainment, or learning progression, so that in their multiple attempts at the Ostensive Demonstration, learners can receive feedback on the level of attainment they have reached, and the ways they could practice before their next attempt. The instrument that administers linguistic knowledge learning should be keyed to be able to score the concepts and terms used in the model of progressive attainment.

On Ostential: It suggests an essential, evidential, and special human performance demonstration. This term currently has not been used in the negative senses of Ostentatious and Ostensibly. There is an unfortunate negative connotation to the word "Ostentatious" that connotes a vulgar display of wealth or beauty, pretention or the like. Also, the word "Ostensibly" is more often than not used to point to possible deceit in a claim someone makes about having attained some valuable goal. One could say that the Olympic Grand Parade of athletes at the beginning of each Olympics is Ostentatious, but remember, in individual sports, only the top 25 in the world in each sport discipline are admitted to the Olympics, and to get there requires many high placements in regional, national, and

World competitions. For team sports, only one team per nation may qualify, and not all nations field a competitive team. The right to be in the Ostentatious grand entrance has been earned multiple times in previous performances of each game. It is *deserved ostentation. Ostential* is a fresh word for this positive display.

The image of the athletes on the medal stands when the gold, silver, and bronze medals are awarded and the national anthem played is obviously very satisfying to the gold medalist, with positive acclaim for all three. There is wide media coverage, and children in the audience may be determined to seek that attainment themselves. It requires very fierce motivation and persistence to go through the years of practice and competing to get to that position. The fourth defining aspect of an Ostensive Demonstration is to keep possible goals and aspirations in mind for the recruiting and motivating power of socially and economically relevant goals. This emphasis on *positive consequences and values* is in strong contrast to the concern in testing to guard against *negative consequences* and have validation evidence strong enough to resist legal claims of unfairness against the test user or maker.

The Olympics is surely a most informative instance of a worldwide recruiting and training system for world-class athletes. Its appeal is wide, and most nations love to have athletes and teams that qualify. The recruiting extends down into school and college sports, and then the best of the best continue to train with a very special educational system—individual coaches and competitive events where observations of others, and sharing, can occur. The rewards are high, and different levels of success or involvement lead to a variety of careers and multiple sources of income. The whole enterprise is made possible by standardized measurement venues and implementation models that preserve the appearance of valid and comparable measurement from year to year and venue to venue. Threats to this perception from practices like doping or compromised judging must be rooted out and punished.

Schools as currently organized do play host to Ostensive Demonstration pathways for training and recruiting for performance domains—athletes, actors, musicians, artists, and to some extent speakers, debaters, cyber security whizzes, and organizers. There is a separate kind of scoring and feedback for these often extra-curricular domains of Ostential Performance. The stand-outs come to the attention of recruiters for post-secondary learning and on-the-job training opportunities.

The main part of the K-12 school is not organized around future work domains but around covering state-mandated standards of academic content and assessing its mastery. The conventional ways of testing used in much of education are decidedly nonuser-centered. The students and teachers are not the buyers of the assessment product, so the product was not designed to their needs. It leads to abhorrence, anxiety, and reduced validity. The big testing days are mandated by district or state agencies or university calendars, not by what is best for any teacher's class or most individuals in it. Many reform-minded educators see this discrepancy between performance domains and academic domains. One adjustment is to use more essay tests and performance demonstrations in academic domains, and they are used increasingly up through the graduate school level. But even with high aspirations to make current systems learner-centered, there is much appealing talk but not an action because the schools are not organized to accomplish it.

Rather than a critique of current testing practices, we instead wish to highlight psychometric and related careers in situations where instrument development and use are likely to have the greatest growth in the marketplace. This growth will be due to attracting end users rather than repelling them; also, by recognizing the goals and aspirations of serious learners. The next section is to explain more about where this user-centered market is going in educational measurement. It also illustrates how the roles of teachers and evaluators are differentiated very differently than in conventional post-secondary institutions. This is a vital lesson in future test use where the learner will be at the center.

8.7.4 Ostensive Demonstrations Entail Reorganized Teaching and Assessment Roles

This section explains a very new, suddenly very large competency-based university in the United States, Western Governors University. It began operations and student admissions in 1997. It had one graduate in 2000 and passed 290,000 graduates in the summer of 2022. It requires that new roles for mentors, evaluators, and others be added to higher education's current transient teaching roles and that assessments are both Ostensive Demonstrations of competency and evidence of content knowledge (based on objective exams). WGU employs thousands of PhD and Masters' level evaluators to score and provide feedback to students and their mentors from performance evaluations.

We may first consider reorganizing for repeatable Ostensive Demonstration. The reader may think that learner-centered assessment can thrive in institutions as traditionally organized. Experience shows this is not done easily. Western Governors University (WGU) provides an example of what happens when a new institution is *intentionally designed* to be learner-centered and competency-based. The Provost and Chief Academic Officer, Marni Baker-Stein explains its student-centered foundation in the 2020 annual report (WGU, 2020):

> If postsecondary education is to realize its promise of creating pathways to opportunity for all, it must commit to a fundamental redesign. We must build with learners at the center—and that means all learners ... *Student-Centered by Design:* Every decision at WGU is based on student success. It's the core of our culture, and it's also built into our business model and organizational structure. Whether academic or administrative, student-facing or behind-the-scenes, every role at WGU is optimized to serve students and every job includes the requirement of being obsessed with student success.

The first graduate illustrates the effect of individual mentoring and student-centeredness. Her name was Jennie Kirch and she graduated on 1 December 2000. A teacher in Roy, Utah, she demonstrated her competency in all required domains for a Masters' Degree in Learning and Technology. The degree was awarded from the WGU Teachers College. After 2+ years of work, she confirmed the student-centered nature with this comment in the 20th-anniversary publication (WGU, 2017, p. 45):

> The entire time I was in school, I felt that everything at WGU was solely dedicated to the student. My mentor, Vince Shrader, was my guide from the get-go. He knew what was needed to be successful in the program and he looked at the skills I had, then put the two together to move forward.

8.7.5 Outline for a Possible Ostensive Demonstration Metadomain for Human Sciences

Mentors are truly the secret to understanding and helping the student over a personalized system where passing off competencies students may already possess is part of the process. At WGU, no lock-step curriculum exists. All courses are online, have frequent starting points, and are optional. The grades in the courses do not matter. WGU was designed to demonstrate an alternative to conventional practices where courses are only available at certain times and places. It is vital for the personal mentors to know each student well and stick with them until they succeed with demonstrated and documented competence for their credentials.

There are teachers, subject-matter tutors, and mentors for particular subjects. These teachers have transient contact with the students leading up to their attempts at periodic competency evaluations.

The traditional teacher's role in higher education is unbundled at WGU in more ways than having personal and subject-specific mentors. The assessment process is taken out of the hands of the

transient individual teachers the student will encounter. The assessment program uses consulting faculty advisory councils to specify and approve the domain contents and the competencies in each subject-matter domain. The tests are created both in-house and by contract to qualified test developers. A large group of evaluators is working constantly to grade student products. Each evaluator has a PhD or Master's degree in areas relevant to the four colleges of Teachers, Business, Information Technology, and sub-MD Medical Professions. The evaluators work at a pace that fits the students' schedules. These evaluators are poised to score the student's written products and media graded on carefully designed rubrics to assure a firm definition of competency. Students can retry more than once a competency assessment they did not pass. The score report includes guidance on what aspects of the competency are in need of more work by each student who did not pass on an attempt. These reports are back to the student and mentor within 72 hours.

As the reader will see from this description, the implementation of Ostensive Definitions and Ostensive Demonstrations along the lines outlined above is possible, but its new functions require reorganized structures. As WGU Provost, Dr. Baker-Stein explained: *"... postsecondary education ... must commit to a fundamental redesign."*

What are the results? The results are remarkable. Evaluation of WGU's impact on student success is found at the end of Chapter 9, Epilogue. The evaluation studies show wide improvement margins for WGU graduates above comparable conventional education in various aspects of success. These results make student-centered and competency-based learning and assessment highly competitive and attractive. It justifies the vision of the 19 Western Governors who sought a better way to get mobility for both underserved students and others in their states—an alternative to building new campuses. It also suggests growth in new opportunities for psychometricians seeking rewarding work.

8.8 Summary and Concluding Thoughts

The story of the SI units is the story of crucial advancements in science. The free and open sharing and coordination of the best science-based units results in great economic and social progress. This is validation that when scientific Metrology is sound, rigorous applied industrial Metrology prospers, and legal Metrology can assist in minimizing the damage from both lawbreakers and unwise laws. This kind of international cooperation bringing social and economic benefit is enough to promote enthusiastic attention in a world where conflict and refusal to cooperate is what gets the headlines (Daston, 2023). It is also enough to envision the benefits that can come to Educational and Psychological Measurement, and other human sciences, as they bring their best and learn from and follow the example of scientific, applied, and legal Metrology in a cooperative world. In doing so we will expand the scope and impact of all Metrology, as human sciences bring understanding to all the STEM and non-STEM human scientists who cooperatively bring about the socioeconomic benefits.

8.8.1 Lack of International Metrology Citations in Human Sciences Measurement Work

Enthusiastic attention cannot be confused with being starry-eyed about the difficulty of the path ahead. We have noted above the lack of citation of international Metrology in the reference books we have been using in Psychology and Educational Measurement. Erroneous conclusions are found in our literature. They are clarified and corrected when the readily available documentation demonstrates quickly how the SI units came to be and how disciplined they are in application. Moreover, by this lack of citation, the examples of international Metrology are not being held as a comparison standard to challenge human science fields to adopt better Metrology practices. Such writers failed to cite

information that is now readily available. In the pre-internet days of many of these writers, it was not readily available. In our day there is no excuse for Psychometricians, Kinanthropometrists, or Metrologists by other names to not be familiar with international Metrology—to understand its role in creating cooperation, fair trade, and peaceful progress among nations. This is the big picture. The smaller picture is that there is a great benefit to becoming familiar with the VIM and the GUM, to understand how units can be traced, and their accurate and fair use controlled, in the commerce of instrument sales.

8.8.2 Looking Ahead: We Will Not Sit Around Waiting

Being enthusiastic about the stunning existence of wide international cooperation in any domain, and finding it in metrology, is a good thing. Being enthusiastic about the very strong connection of this cooperation in sharing common measurement units that lead to social and economic progress is another good thing. This does not mean we believe the human sciences will quickly benefit from the social and economic boon of acceptance in an international community. That may come slowly. The long, slow history of acceptance of units in the STEM disciplines is sobering. The definition of the meter and kilogram occurred 250 years ago. It would be unwise to wait around for the physicists and chemists in particular to relent on the fundamental necessity for quantities and invite us in with our numerical values that do not even support ratios! How can this pure and beautiful edifice of the international system of units permit a lot of questionable new measures that cannot now, or in any currently conceivable future, mesh into the coherent system of units? It is not so bad having very few that cannot be coherent, like hardness. Moreover, how can the uncertainty of numerical values assigned to properties of objects (including human beings) be trusted as measures? The VIM 3 (JGCM, 2013) contains a brief preliminary paragraph that indicates the JGCM's expectations of which professions they serve:

> ... This Vocabulary is meant to be a common reference for scientists and engineers, including physicists, chemists, medical scientists, as well as for both teachers and practitioners ...

Since waiting around is pointless, we can regard the coming debates as a form of entertainment. The acceptance of the new definitions proposed by Mari, Wilson, and Maul is provoking a lot of discussion—of an unpredictable duration. In addition to being in the audience for the debates, so far as they are made available, we might consider participating in promising beginnings that could lead to widely shared human science units within pioneering nations. An example of how efforts in the United States (with participants from Canada and Australia) might lead to something important is the Open Skills Network, a nonprofit organization found at https:/www.openskillsnetwork.org.

8.8.3 Open Skills Network and Rich Skills Descriptions

One of the founding organizations in the creation of Open Skills Network, WGU has made available for members of OSN 13,000 rich skills descriptions (RSD). The OSN does not collect and share libraries of skill *names* but of *Rich Skill Descriptors*. From the website:

> *Rich Skill Descriptors (RSDs) are machine-readable, searchable data that include the context behind a skill, giving users a common definition for a particular skill and help to make it understandable and transferable across the learning-earning landscape.*

Developing a common language of skills that communicates across educators, learners, assessment developers, and employers creates a community of providers and users. This common language needs to be understood by the educators who are preparing the learners for certificates or degrees (and assuring by assessment that they are competent). Of course, the learners need to understand what competencies they could repeatedly demonstrate with a credential in their portfolio of capabilities. They need to be articulate about the nature of each skill and about the evidence they have that they possess it. Many employers are signing up for OSN, seeking job candidates with credible evidence that they have the skills required for the job. Even without having been inspired by international Metrology, it seems that the Open Skills Network might have many of the values that led to the wide use of common weights and measures in Metrology, but Rich Skills Descriptors are in the Ostensive meta-domain of human competencies.

RSDs are not common measures. They only *lead to* a more deeply defined account of the stages of progression for a skill, as discussed in Chapter 5 of this book. For a particular job, a particular RSD is already domain-specific, but some skills obviously generalize across multiple jobs in multiple domains and similar skills in different but fraternal domains will be found. Model-based reasoning is possible for each skill if it has been developed by domain analysis and model development, and the Person model gives the levels of progression, the task model and evidence model of how to assess it. Don't forget that Evidence-Centered Design by its very nature is a design language for the whole process of assessment development. If you read the associated literature, you will find that *design patterns* are also reusable.[7] A technology of assessment development is getting more tools and more standardized parts. The field of Industrial Metrology for assessment development is emerging. Measurement instruments for all or most of these RSDs will be developed.

What the Open Skills Network provides is a common language so that every good assessment development organization can offer instruments, and either play the proprietary game that testing has so far played, or start down the bold path of *sharing widely* the core definitions and later the stages of progression discovered about many skills. A prudent guess would be that scientific discoveries about the progressive nature of each skill—knowledge that might greatly accelerate learning progress—will stay within the discovering institution's educational program for a least a time long enough to assure that the model of progressive attainment (or ECD Student Model) is well validated.

8.8.4 Putting Rich Skills Sharing "in Context," a Comparison to the History of SI Units

The Open Skills Network received substantial start-up funding from Walmart, who is an eager beneficiary of more and better-qualified job candidates with the skills they need. The US Department of Labor has assisted in conferences getting it launched. *Many* additional employers are signing on, and the US Air Force has a funded project for their domains of important skills. It is interesting how much buzz and sharing of a new concept comes when it resembles the essential nature of Metrology (whether Scientific, Applied, or Legal). It is open, it is shared, and it has quickly recognizable value to employees, employers, and educators who wish to remain competitive for students. It makes both local instrument development and industrial development more attractive. Each teaching

7 PADI technical report #1 Mislevy et al. (2007) is about Design Patterns for higher levels of capability than is thought of in using the language of skills—even RSD's. But consider this: scientific thinking like model-based reasoning may require learning progressions over more than one domain. A complex Design Pattern is developed and can be shared for future projects. Design patterns for kinds of RSD's will be the same—Shared Design Patterns as well as shared RSD's.

organization with developers and many instrument development companies can tap into the growing resources of the OSN.

If the open sharing of Rich Skills Descriptions is analogous to what it was that led to the wide sharing of SI units, it will continue growing rapidly, as WGU has so far done. The Ostensive Metadomain may be productive in creating significant social and economic capital, with less proportionate income for item development and more for the mechanisms of performance demonstration and judging that more readily produce valuable feedback and coaching information.

8.8.5 As a Send-Off

As graduates in quantitative human science fields, you may have to work sometimes in "common-law" testing programs for conventional high-stakes testing programs that do not delight or empower the test takers. Consider the evidence of social impact and economic growth we have tried to demonstrate by going into new ways of thinking about measurement and testing. Go with the greatest Scientific Metrology and the most test-taker-friendly applications you can find. There are more jobs in the applied or industrial side, but be involved or in touch with the ground floor, that is, in Scientific Metrology, as all of the authors of this book recommend. However, keep trying to understand the Industrial and Legal Metrology forces that put the new measurement instruments productively and profitably into the *Information World* of interpretation and use.

Acknowledgments

While the Mari et al. book came to our attention in May 2022, prior knowledge leading to the early drafts of this chapter came about by an unexpected encounter. A successor author, Victor Bunderson, shifted in late 2018 from being a Foreword–Preface writer to being a coauthor after his time freed up. He and his wife were enjoying a 4-week guided tour of Thailand, Cambodia, Laos, and Vietnam in Fall of 2019. He did a small amount of work on drafts of Chapters 1 and 3 some evenings while relaxing in hotels. On this same tour bus was Dr. Yi-hua Tang, a physicist-metrologist. He was taking superior pictures and enjoying time off he had not had for years. He had retired as head of the Voltage Lab at the NIST, Gaithersburg, Md. He supported the needs for voltage measurement in Watt Balance and continuing improvements. The work from the NIST Voltage Lab was averaged (weighted by uncertainty of each lab) with similar estimates made across several international Metrology institutes. It led to the retirement of the IPK, better estimates for the Ampere, and also better estimates of the charge of the electron as discussed above in reference to International Metrology Day in May 2019 (Section 8.3.2).

On buses and in cafés across four countries, the two conversed. Tang spoke to Victor about the English graduate student Brian Josephson and his idea that later led to a Nobel Prize in Physics. He mentioned how the Josephson effect led to the measurement of the elementary charge of a single electron which is most precisely measured in terms of the Josephson constant and the Von Klitzing constant. That constant is related to the quantum Hall effect.

He explained to Victor how the Kibble Balance worked. At several times he led Victor to downloadable references at NIST and BIPD. Victor was struggling with sections where Petr Blahuš was reasoning about how measurement in the hard sciences took place, using the examples of length, mass, and temperature. Dr. Tang introduced Victor to official documentaion about these specific units. Metrology has a tradition of openness and sharing across nations. For example, the SI Brochure gives in its appendix brief histories of how the units evolved as the underlying science improved. It tells when each unit was accepted internationally. Other references give the structure

of the entire Scientific, Applied, and Legal Metrology enterprise that is required to produce and distribute measurement instruments that faithfully measure the units.

We also have benefitted greatly from the Mari et al. book *Measurement Across the Sciences*. Section 8.5 summarizes the substantial impact that Luca Mari, Mark Wilson, and Andrew Maul had upon our efforts to weave Metrology into this book in far less than a year after we discovered it. Mark Wilson was also co-editor of the book *Advances in Rasch Measurement, Volume 1* (2010). These editors included William Fisher, Jr., another friend and international metrology advocate. They took several papers Victor's research team had presented in 2002 on extensions of Messick's ideas about Domain Theory and helped the authors refine them with excellent editorial guidance. This work was particularly instrumental in writing Chapters 4 and 5 of this book.

8.9 Study Questions

Forty Discussion Questions

Q1 What are the three major branches of the international organizations of Metrology and what is the purpose of each? (Section 8.1, first paragraph)

Q2 What are the four avowed purposes of this Metrology chapter? (Section 8.1, second and third 3rd paragraphs)

Q3 What does the STEM acronym represent? What are examples of STEM disciplines? (Section 8.1 and footnote 1)

Q4 When was the first Metrology International Treaty Convention held? What were some of the outcomes? (Section 8.3.1, first paragraph)

Q5 What are some sources to guide 21st century development in measuring human cognitive, conative, and affective performance that could lead to Metrologically verifiable units? (Section Section 8.1, fifth paragraph)

Q6 Who were the two astronomers chosen to make a more accurate estimate of the measure of the quarter circumference of the earth passing through Paris? (Section 8.2.1, fifth paragraph)

Q7 Which two members of the mathematics team in this quest were opponents to one another in a priority dispute over who discovered the least squares method? (Section 8.2, end of fifth paragraph)

Q8 The physicists of the Ferguson Committee held that _____ was a necessary precondition, a demarcation criterion, for true scientific measurement. (Section 8.2.1, paragraphs 1 to 3)

Q9 Relate the story of S.S. Stevens publishing a blockbuster paper in *Science* in 1946 as a rebuttal to the Ferguson Committee's critique of his Sone scale. (Section 8.2.1, paragraphs 1 to 3)

Q10 Relate also the story of Percy Bridgman's objection to psychologists "crediting" him with being the originator of their favorite irrational strategy of defining variables using so-called "operational definitions." (Section 8.2.1, paragraphs 6 & 7)

Q11 Why is Metrology almost never discussed nor mentioned in the psychometric literature? (Section 8.2.2)

Q12 What motivates new instrument development in Scientific Metrology? (Section 8.1, paragraphs 1 & 2)

Q13 What are some of the activities of Applied Metrologists? (Section 8.1, and Section 8.4)

Q14 What are some of the activities, issues, and concerns in Legal Metrology? (Section 8.2)

Q15 How and when did the International Bureau of Weights and Measures (BIPM) come into existence? (Section 8.1, paragraphs 6–8)

Q16 What are the tasks and objectives of the BIPM? (Section 8.3.1, first paragraph)

Q17 How are the BIPM, the CIPM, and the CGPM functionally related to one another? (Section 8.3.1, paragraphs 1 & 2)

Q18 What kinds of advances have come from progress in the accuracy and scope of Metrological measurement? (Section 8.3.1)

Q19 What is the name and acronym of the Metrological Institute of the United States? (Section 8.3.1, paragraphs 6–9)

Q20 What have been some of the achievements of NIST over the past century? (Section 8.3.1, paragraph 6–8)

Q21 What is unusual about the Watt Lab in Gaithersburg, Maryland? (Section 8.3.1, paragraph 7)

Q22 What advances in measurement have come about because of the Kibble Balance? (Section 8.3.1, paragraph 8)

Q23 Why was International Metrology Day, 20 May 2019, so significant? (Section 8.3.2)

Q24 Why was it necessary to replace physical artifacts with fundamental constants of nature? (Section 8.3.3)

Q25 What is the rationale of the Manifesto for establishing International Metrology in ***all*** of the sciences? (Section 8.5)

Q26 What justification can be given for extending the VIM to the human sciences? (Section 8.5)

Q27 What special problems and concerns are there for measurement in the human sciences? (Sections 8.5.2 and 8.5.3)

Q28 Summarize Richard Brown's description of Metrology as being "at the level of metathought above measurement"? (Section 8.6)

Q29 How are Extensive and Intensive collections of properties "metadomains"? (Section 8.6)

Q30 Discuss the author's claim that the small non-profit group Open Skills Network, in only three countries could be a forerunner to shared scales, if not units, in educational measurement[8].

References

Alder, K. (2003). *The measure of all things: The seven-year Odyssey and Hidden error that transformed the world*. Free Press.

Brown, R. J. C. (2021). *Measuring measurement - what is metrology and why does it matter? Measurement*. (London), 168: 108408. https://doi.org/101016/j.measurement.2020.108408.

Bureau International des Poids et Mesures (1998). *National and international needs relating to metrology: International collborations and the role of the BIPM*. Intergovernmental Organization of the Convention of the Metre.

Campbell, N. R. (1920). *Physics: The elements*. Cambridge: Cambridge University Press. Reprinted (1957) as *Foundations of Science: The Philosophy and Theory of Experiment*. New York: Dover.

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London, Longmans.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, 3, 186–190.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Lawrence Erlbaum, Mahway, NJ.

Daston, L. (2023). Being in the room together. *Annals of the New York Academy of Science*, 1–3. doi: https://doi.org/10.1111/nyas.15077.

De Bièvre, P. (2012). The 2012 International Vocabulary of Metrology: "VIM". *Chemistry International—Newsmagazine for IUPAC*, 34, 26–27.

International Bureau of Weights and Measures, & International Organization for Standardization (1993). *Guide to the expression of uncertainty in measurement* (1st ed. 1993). International Organization for Standardization.

IUPAC (2007). *Quantities, units and symbols in physical chemistry*, 3rd ed. Research Triangle Park, NC: International Union of Pure and Applied Chemistry, Physical and Biophysical Chemistry Division.

Jorio, A, & Dresslhaus, M.D. (2016). *Nanostructured materials: Metrology in material science and materials engineering, reference module in materials science and materials engineering*, Science Direct. (accessed Nov 2020)

JCGM 200:2012 International vocabulary of metrology - Basic and general concepts and associated terms (VIM) 3rd edition.

Josephson, B. D. (1962). Possible new effects in supercondctive tunnelling. *Physics Letters*, 1(7), 251–253.

Koch, S. (1992). Psychology's Bridgman vs Bridgman's Bridgman: An essay in reconstruction. *Theory & Psychology*, 2, 261–290.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Volume I: Additive and polynomal representations*. Academic Press.

Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions and applications*. Information Age Publishing.

8 BIPM paragraph (second page of Section 8.1).

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the sciences: Developing a shared concept system for measurement*. Springer Series in Measurement Science and Technology. Springer.

Mari, L., Wilson, M., & Maul, A. (2023). *Measurement across the sciences: Developing a shared concept system for measurement, 2nd Edition, e-edition is Open Access*. Springer, e-ISSN 2198-7815.

Michell, J. (2008). Is psychometrics pathological science? *Measurement*, 6, 7–24.

Mislevy, R. J. (2009). *Validity from the perspective of model-based reasoning*. CRESST Report 752. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mislevy, R. J. (2018). Sociocognitive Foundations of Educational Measurement, Routledge, New York, NY 10017.

Narens, L. (2002). *Theories of meaningfulness*. Mahwah, NJ: Lawrence Erlbaum Associates.

Steiner, R., Newell, D., & Williams, E. (2005). Details of the 1998 watt balance experiment determining the Planck constant. *Journal of Research of the National Institute of Standards and Technology*, 110(1), 1–26.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103 (2684), 677–680.

Stock, M., Davis, R., de Mirandés, E., & Milton, M. J. (2019). The revision of the SI—the result of three decades of progress in metrology. *Metrologia*, 56(2), 022001.

Tang, Y., Seifert, F, and Schlamminger, S. (2016) *Improvement of leakage resistance of PJVS for NIST-4 Watt balance experiment*. Ottawa: CPEM 2016 Conference Digest, [online] https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=920019 (accessed Feb 3–4, 2023)

9

Epilogue

Aspects of This Psychometrics Book

CHAPTER MENU

9.1 Improbable and Unusual, 347
9.1.1 Unusual—The Metrological Slant This Book Takes, 348
9.2 Unusual Scientific and Pedagogical Perspectives, 348
9.2.1 An Unusual Scientific Perspective, 348
9.2.2 Unusual Content and Structure, 349
9.2.3 Mathematical Appendices 1–5, 349
9.2.4 Software Computational Guides for Practice in Using Psychometric Tools, 350
9.2.5 Benefits of the Long Preparation for the Book, 351
9.3 The Value of Design and Domain Theory, 351
9.3.1 Domain-Specific Analysis and Modeling: Their Use in Assessment, 351
9.3.2 A New Metadomain Using Ostensive Definition and Ostensive Demonstration, 352
9.4 New Approaches to Measurement, Validation, and Educational Evaluation, 353
9.4.1 A Progressive Approach to Measurement, 353
9.4.2 Future Perspectives on Validity and Validation, 353
9.4.3 The Western Governors University Experiment and the Ostensive Demonstration of Skills, 355
9.5 Summary, 357
References, 358

9.1 Improbable and Unusual

At the end of this improbable and unusual book, we try to give some perspective on the benefits of it being both improbable and unusual. It would not have existed had Petr Blahuš not visited Brigham Young University (BYU) in 2002–2003, where he added materials to his manuscript and obtained permission to use the CERAD battery with its verbal ability and executive ability latent factor materials. Nor would it exist if Bruce and Susan Brown had not visited Petr in Prague in 2010 when his health was rapidly declining, and made a plan for the completion of Petr's fine psychometrics manuscript and to arrange for its publication in English.

The four authors of this book (Blahuš, Brown, Bunderson, and Olsen) met together for only one day in 2002 and had no idea that two of them (Brown and Olsen) would toil together for 7 years, joined by the third (Bunderson) for five of those years, to complete the book. Such persistence took a sense of kinship with the senior author, which truly exists. We respect him as a friend, an international scholar of note, and as a wise and insightful scientist.

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

9.1.1 Unusual—The Metrological Slant This Book Takes

It originated from an improbable 2019 meeting in Thailand between continuing author Victor Bunderson and Dr. Yi-hua Tang, a noted metrologist, newly retired as head of the Voltage Lab at NIST. This resulted in a draft of Chapter 8 for this book, encouraging measurement experts in the human sciences to pay close attention to the edifice of international measurement: scientific, applied, and legal. If this had not taken place, the continuing authors would not have seen quickly the importance of the book by Mari, Wilson, and Maul (2021), which introduces new definitions of direct and indirect measurement phrased in the definitions of the *Vocabulary International of Measurement*, or "VIM." While the familiar name VIM is used for brevity, the correct citation involves the Joint Committee on Guides for Metrology, or JCGM. See the references for this chapter. VIM 3 is the definitional source for scientific metrology in physics, chemistry, engineering, and medicine. The new definitions in this important book open up the restrictions VIM imposes on measurement terminology and possibilities so that it can be used by all sciences. The international metrology slant of this book is central to helping the human sciences rethink how our measurement enterprise is evolving.

9.2 Unusual Scientific and Pedagogical Perspectives

9.2.1 An Unusual Scientific Perspective

Petr Blahuš's perspective as a Kinanthropometrist is seen by his continuing coauthors as being uniquely conducive to realist explanations of human constructs, whether visible athletic movements or invisible thinking processes. There had to be a feeling that the book could possibly become quite useful to our colleagues and that it brought important things together in an uncommon manner to make it worth the long, difficult process.

The kinanthropometrics and sports science perspective, just one part of Petr's rich background of experience, is part of this hope for major progress for measurement in the human sciences. Bodily movements are more clearly visible and thus easier to see and understand as real sequential processes in any standardized performance, while thoughts and mental processes are mainly observable through their interaction with tasks or ratings of performance demonstrations by self or observers. We also see in his domain of sports science a metaphor for how domains of human activity use standardized measurement instruments and scoring transducers and/or judges to create industries like the Olympics or professional sports. Consider the Olympics: The argument of comparable, even invariant measurement between the impressive "instruments" in Olympic venues enables such industries to exist at all. It enables records to be bested at one time and at venues quite different from one another. It enables ranking of gold, silver, and bronze medalists.

The cover of this book is inspired by the Olympics metaphor, mentioned in several places in this book. In Chapter 5, high jumping was used to illustrate aspects of computer-adaptive testing. The cover picture depicts an athlete practicing the Fosbury Flop in preparation for high jumping competition, in a bright new stadium with no top-row observers. Such practice is needed when athletes arrive at the city sponsoring the Olympics, and soon they are able to practice in the actual venue where they will later perform. Repeated practice of ostensive demonstrations of expertise like this, followed by wise coaching feedback, is a theme of this book. We recommend it be widened to many more subjects than sports. Not shown in the cover photo are the coaches who accompany the athletes. The expert coach, who usually has been an athlete in the sport previously, would have given advice following the depicted performance. This advice may have dealt with details of the athlete's

approach, take-off, clearing the bar, or landing. The Fosbury Flop is an example of how good science can transform sports practices. Fosbury was an engineer who understood that his center of gravity could be many centimeters lower than with techniques used by his competitors. After his gold medal, the Flop replaced all competing methods over a period of years as others learned it. Not pictured also is the mind coach, which many athletes are provided through their sponsoring nation or corporate sponsors. Anxiety and attendant lack of focus have thwarted many athletes who succeed at lower levels but get rattled at the Olympics.

Another benefit of the Olympic metaphor is that international sports provide evidence of the importance of domain-specific *human values and consequences*, a disputed part of the validity concept that is indisputable in the Olympics. People in the domain are attracted to these positive values and consequences, which leads to recruiting chains and eager engagement in difficult practice with feedback from excellent coaches at the high end of criterion scores. The educational measurement dispute is about instrument developers being responsible for negative consequences (such as unfair intelligence testing of new immigrants), whereas it is in fact the positive consequences and values that make the domain socially and economically vibrant. Positive values and consequences cannot be ignored by instrument developers. Legal threats are useful goads in inducing instrument developers to spend time and effort validating their claims and protecting against negative consequences. But it is the positive consequences and the criterion examples that make a domain exciting for its participants and the family and friends who encourage them.

9.2.2 Unusual Content and Structure

This book is also unusual in the mixture of older and newer content and styles. The first author did not use chapters, but sections that he tied together with copious use of backward and forward references. This is not student-friendly. As his chosen quotes at the beginning of the Preface proclaim, he is strongly mathematical. This reduces the ease of reading. The two initial continuing authors (Brown and Olsen) desired something usable by their own students and wanted to make the book as learner-friendly as possible. The third continuing author (Bunderson) had before retirement from BYU been teaching in *Instructional Psychology and Technology*, a field that studies design models, methods, and theory for both assessment design and instructional design. All three are deeply committed to learner-friendly assessment *and* the instruction that should precede and follow it. In his teaching at BYU Petr demonstrated these values, and in his lifelong career in Prague he lived and worked among strong demonstrations in sports teaching and coaching. You have seen by now what appendices, computational guides, and synthesizing explanations are placed around the sections in each chapter to offer additional student-friendly assists that Petr did not have time to add.

9.2.3 Mathematical Appendices 1–5

The five mathematical appendices use simple examples of the math behind central tendency, variance and covariance, correlation, matrix algebra, and bivariate and multiple regression. Results may be calculated by hand while seeing the mathematical notation, the graphical representation, the matrix algebra representation, and the procedural steps. These multiple representational perspectives provide a Rosetta stone for a variety of representations of the same concepts, in order to promote strong learning for various types of learners. Also in one case, Appendix 5 on multiple regression, Stata computational procedures are demonstrated. These appendices are learner-tested and have quizzes at the end.

9.2.4 Software Computational Guides for Practice in Using Psychometric Tools

Frequently emphasized in the chapters that require them, individual practice with "the tools of the trade" are as valuable as the best pages of text. The Computational Guides are built around readily available software packages. This is a sign of how different our times are from the days of Blahuš and McDonald. It is becoming increasingly difficult to obtain useful copies of the pioneering software packages such as COSAN, TestComp, and the special programs McDonald developed with Colin Fraser, that will run on today's planned obsolescence computers. However, there exists a burgeoning academic software industry in statistical methods, mathematics, and psychometrics. A number of colleges and universities have access agreements with the particular package developers we selected for student use. But more can be said about selection of, or development of, one's own software when you become a graduated professional.

As has been discussed throughout the main text and in the statistical appendices and accompanying computational guides, psychometric practice has benefited from important foundational work and recent developments in psychometric modeling, including methods for classical test theory, item response theory, and factor analysis. The latent common factors of factor analysis and the latent traits of item response theory help us address scientific constructs of interest in the social, behavioral, educational, and other human sciences. These constructs may be ideally represented in the latent structure of data from multiple observed indicators using well-designed psychometric instruments.

To oversimplify somewhat, where the observed indicators are quantitative or metric, we usually apply traditional linear factor analysis methods, and where we are working with binary or categorical indicators, we generally utilize (sometimes multidimensional) item response theory methods. So we might say that at some level, both latent common factors and latent traits can be treated as the same kind of latent variable where any difference is simply in the nature of the observed data being modeled. Latent variable methods can be used to evaluate traditional assumptions and models of classical test theory and obtain reliability estimates and other classical test theory results.

Generalized latent variable modeling software such as Mplus, gsem in Stata, and lavaan in R implement a broad statistical modeling framework that can encompass both traditional factor analysis and item response theory modeling. Mplus and lavaan provide both confirmatory and exploratory analysis of continuous and categorical items. Mplus, gsem, and lavaan allow for multiple group analyses and effects of external predictor variables (or covariates), facilitating measurement invariance testing and differential item functioning analyses, along with the estimation of structural equation and explanatory item response theory models. These programs also provide ways to address the special challenges and opportunities of longitudinal and multilevel or clustered psychometric data.

The wide availability of MPlus and Stata, plus the opportunity to demonstrate important features with smaller examples that fit the free version of MPlus made these two good choices for the Computational Guides featured in this book. They provide familiarity that encourages growth into other more advanced features available in each package but not practiced in the Guides. In professional work, it may be necessary to create new approaches within familiar software packages, but sometimes new methods are necessary. If they can be expressed in matrix algebra, one can use the corresponding functions in available packages (such as R, SAS, Stata, and even Excel) to define new processes. Sometimes new software for specialized research must be developed as in the next paragraph.

McDonald (1999) proposed a unified way of thinking about psychometrics by using ... *a general nonlinear common factor model which includes item response models as special cases, and includes the (linear) common factor model as an approximation.* Similarly, we are seeing the interplay between latent factors and latent variable models leading to a unified statistical modeling

framework for doing psychometrics. Software tools are becoming more capable of supporting this work. A dedicated classical test theory or item response theory or factor analysis software package can be more user-friendly and convenient, whereas particularly specialized software may be more readily customized or provide a richer output repertoire. Even commercial and freely available general-purpose statistical software is beginning to incorporate more features useful for psychometrics. Nevertheless, general latent variable modeling software is probably the most useful for learning and performing basic and extended psychometric analyses. More advanced or specialized work may require R packages such as those described in Choi and Asilkalkan (2019) or custom tools built with Python, Stan, BUGS, JAGS, etc.

9.2.5 Benefits of the Long Preparation for the Book

For various reasons noted in the preface, publication of this book has taken a number of years. The resulting book thus has the unique feature of juxtaposing Petr's key content written 10 to 15 years ago with game-changing developments since then. This highlights the importance of the innovations, which are not as easily seen in contrast over only a few years. Moreover, for graduating psychometricians and related scientists in the human science domains, as you go out into the world of 2025 and beyond, you will observe many practices *still in use* that Petr as a keen observer deplored and sought to replace. There is a lag of decades in dissemination of new ideas and the difficult changes they require. In Chapter 4, we quoted ACT-testing expert Cynthia Board Schmeiser who comments on certain unproductive *common law* practices that have crept into psychometric work, "... a common law of test development conveyed through advice and opinion throughout its history." She wrote this in the first volume of the *Handbook of Test Development* (Downing & Haladyna, 2006, p. 760), which connects with the Joint Standards presumably as an attempt to counter the growth of these "common-law" practices. Even the most odious ones, that Petr so deplored, are hard to root out.

Petr was unable to do much student testing as the manuscript grew, and we discovered that some of his sections were written at a high graduate level. A benefit is that it provides a reference source for a mathematical approach that does not rely on proofs, which you as a professional may find useful as your own mathematical grasp increases over the years. Providing scaffolding for students below this level required substantial additional materials to build lower concepts and skills and also a chapter structure with a wrap-around to group key topics. Since our delay in getting Petr's text completed has a double edge, consider the next few positive developments.

9.3 The Value of Design and Domain Theory

9.3.1 Domain-Specific Analysis and Modeling: Their Use in Assessment

Petr speaks positively about design. Some of Robert Mislevy and colleagues' writings on evidence-centered design (ECD) were available to him, but he did not cite them or have them specifically in his references. He is not alone in this. As Herbert Simon (1996) reminds us in his *Sciences of the Artificial*, engineering schools had abandoned the "methods of the cookbook" that flourished for a while until a strong reaction ejected them. McDonald (1999) used the "cookbook" term derisively and as something to be avoided. Perhaps Petr saw some writings on design in domains like curriculum and educational psychology in this light. After pointing out the reaction to the cookbook, Simon had to remind his readers from many disciplines that there could be a science of

design. One has to read the Mislevy and Riconscente's (2005) work carefully to realize that it is so comprehensive as to create a design language for scientific evidence-based assessment development. Users keep writing their own special features as they adapt the design process and its language to their specific domains. It is a design language for assessment development that requires domain-specific insights of the user. It is not a method of the cookbook but may be dismissed as such without a deeper look.

In the text, Petr often calls out for the discipline of theory at some stage of developing latent factor models. We have interjected comments about the value of domain analysis followed by model development and the use of model-based reasoning as scientists do in many fields. We have promoted the disciplines and thinking of ECD. There are specific models developed in the second and third layers of assessment design using ECD. They rely on the first layer, domain analysis, which recognizes as Sam Messick did that something he called domain theory was needed for each particular domain. First, the student model is developed; second, the task models; and third, evidence models for claims the instrument is making. Then model-based reasoning can provide the discipline of highly relevant theory that Petr was always seeking. Petr is critical "of so-called "constructs" and rightly so," as constructs seldom start out being domain-specific. Domain models in ECD do start rooted in domain analysis, and provide specifically relevant models that together make possible domain-specific reasoning to guide the decisions at hand. This does not replace the idea of construct validity, but the combination of models is ever so much more relevant and useful than the word "construct." This word is better used as a verb than as a noun during assessment development.

9.3.2 A New Metadomain Using Ostensive Definition and Ostensive Demonstration

As explained in Chapter 8, the Metrology chapter, ostensive considered a metadomain, as are extensive and intensive physical and chemical metadomains of kinds of measurable properties. For some reason, Petr associated ostensive with direct measurement of extensive attributes that sustain a concatenation operation. He wrote of demonstrating concatenation by measuring length or mass. On more than one occasion, he used the term *ostensive concatenation* to identify direct measurement and, by the addition of *ostensive* to concatenation, to make it clear that by the demonstration of laying down the calibrated rods to measure a field or putting the calibrated weights on a balance to measure mass, one could *publicly demonstrate* evidence that the measurement was indeed direct and supported concatenation.

We were puzzled about why he would restrict his options by focusing on extensive properties, the only ones that support concatenation. The huge array of intensive physical and chemical properties does not support it. All he needed to do to generalize ostensivity to a metadomain was to insert one word—demonstration. Call it an ostensive *demonstration* of concatenation, and then let concatenation become M (measurement) and "M" can include any measurable human attribute. As we have explained in Chapter 8, ostensive definition involves pointing to the people who are exemplars of different levels of criterion performance, and ostensive demonstration is the standardized performance situation for which rating scales or other indicators are developed. Such scales are used to obtain evidence of where on the scale the person doing the performance falls. We are grateful that he did obsess that there must be something more here, by bringing *ostensivity* up repeatedly.

But he did enough to stimulate our recognition that ostensive definition and ostensive demonstration fit his many studies of human movement in sports, and also, that it fit the performance demonstrations Western Governors University (WGU) uses to establish competencies at criterion

levels in the many domains within their Teachers College, their Business College, their IT College, and their Health Professions College.

Borsboom (2018) was given the great honor of making a special anniversary presentation to the Psychometric Society on the topic *Philosophy of Psychometrics*. It is well worth the time it takes to click on the following link and watch Denny Borsboom (2018) on video doing what Petr was only describing in print more than a decade earlier: https://vimeo.com/256145513. He gives a very clear and memorable *ostensive demonstration of concatenation*. Borsboom demonstrates to his audience the creation of a 1–1 correspondence between several carefully crafted clay balls representing one, two, three, and five units of weight. He creates the 1 unit, then the 2 and 3 weighing them each time with the balance, then shows that the concatenation of the 2 and 3 unit balls balance a 5 unit ball.

Definition by pointing to vivid examples can be very convincing. Watching athletes set a new world record at an event is very convincing of their competence at the highest criterion level. Ostensive demonstration is convincing. As we mention below, Boorsboom has been placed on the extreme scientific metrology side of the validity debate by Mari et al. (2021). They see him as leaving out the information world side, while Michael Kane does just the reverse. Perhaps each occasionally goes a little too far, but in so doing they clarify and contribute to our dialectic perspective.

9.4 New Approaches to Measurement, Validation, and Educational Evaluation

9.4.1 A Progressive Approach to Measurement

Even though the title, *Psychometrics, Test Theory, and the Latent Factor Model* leaves out the term *measurement*, measurement is clearly at the center of the book. Petr understood it as foundational and tried to innovate with "*weak associative measurement*" and with ostensive definitions and demonstrations. The former has not gained traction in the 15 years or so since he wrote it, but something much better has replaced it. The Mari et al. definition of direct measurement fits perfectly into his model-based scientific arguments. If his concept of ostensive demonstration gains traction, he gets a full kudo. He gets half a kudo for defining the scientific need that is now fulfilled by the direct measurement concept of the metrologists.

We have added a perspective that is not common in psychometric texts, an appreciation of international metrology, scientific, applied, and legal. The timely arrival of Mari et al. (2021) *Measurement Across the Sciences* upstepped what we were able to say and do in this book. It is not only the perfect replacement for Petr's struggles to find a kind of measurement that fits his scientific aspirations, but it can clarify thinking about the role of measurement in the human sciences in ways we have not fully imagined in this book, though we quickly saw the value of *Measurement Across the Sciences*. The new definitions for all sciences for direct and indirect measurement beautifully replace what Petr was trying to do with Weak Associative Measurement. More than that, the perspective of international metrology sheds considerable light on current issues and on possible paths for the growth of professional opportunities for graduates in this field. Consider how it affects unresolved issues in validity.

9.4.2 Future Perspectives on Validity and Validation

This book points out both unnecessarily incompatible sides of the arguments on validity theory that have characterized the last 15–20 years. The mainstream approach, championed by Michael Kane

and many others, is that no mention is made of the importance of the instrument in validity, but validation applies specifically to the interpretations and actions taken on the basis of test scores obtained at a particular testing occasion. Measurement is not mentioned, nor is the instrument. Borsboom, Mellenbergh, and Van Heerden (2004) offered instead that: "A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes." In chapter 4.3.4 of Mari et al. (2021), the Boorsboom definition of validity was found acceptable and was incorporated into the definition of direct measurement. They did not find value in the Messick-Kane approach. They applied the Boorsboom definition to the first process that takes place in their intentionally designed, empirical, and informational 3-stage process of direct measurement. This first process is transduction and if a property exists and causes changes in the instrument, *transduction* can take place and move to the second process of applying the instrument scale. Transduction takes observations in the *empirical world* into an instrument with a scale, and out with a measured value on a publicly accepted scale into the information world. In the VIM, the concept of validity as practiced by the sciences that use it is reduced to mere accuracy of obtaining the value for what is being measured. All the information the world gets is that number, not any interpretations or optional actions.

There is no place for interpretations and actions in scientific metrology. Instrument developers, however, have that responsibility in full. They operate fully in the information world and both for practical marketing reasons and legal accountability reasons need to attend carefully to interpretations and recommendations for actions. Petr points out in his discussion of Table 3.3 that the value produced by the measurement needs *explanations* for what the measured value means—interpretations. Those who administer the test also need to answer the questions of the test taker about what the score means. In the Ostensive Measurement of human properties discussed in Chapter 8, the learner is permitted multiple attempts at passing each ostensive demonstration of some skill or competency. The test taker must quickly afterward receive interpretations of the score received and its relation to passing or failing that competency. Also, without being told a specific correct answer, the learner needs to know what to study or practice next, to be ready for the next attempt, so their next score will probably not fall below the criterion again. Interpretations and advice for actions are the heart and soul of what is valuable in the information world. Moreover, Kane is right in warning that if instrument developers do not have valid evidence for their claims about what the scores mean, they may face legal consequences.

It would appear that the scientific metrology view is only about the measurement, and validity is reduced to accuracy. Contrarily, in the dominant view of validity in Educational Measurement taught by Kane and others, the instrument and measurement are not even a part of validity. Validation is only about interpretations and actions following a testing event.

Let us create a scale for Total Metrology: *scientific* ➔ *applied* ➔ *legal*. These parts are necessarily separate, and each one leads to the next one. Then let us position validity theories along this as if it were a rating scale for judging how much of a theory or practice is relevant to each of the three types of concerns, starting with the scientific.

The position of Mari et al. on this scale is unambiguous. They precisely cover the scientific and do not spread farther. Their definition is fit for the VIM, which is all about scientific units being accepted and used internationally. Mari et al. could be criticized for how inadequate they are going to be in the information world concept they have so helpfully given us. In the information world, they cannot get by with just the numerical value of the calibrated measure produced by the instrument. Developing instruments is clearly in the applied metrology domain, and laws are written governing measurement instruments and their uses. Interpretations and possible actions are the highest values in the pragmatic world of economically and socially sustaining a measurement instrument development activity.

Consider Michael Kane's position on validation. It is centered on the applied and legal aspects of metrology and is less attentive to the scientific. His writings often appear narrow, yet sometimes spread to the scientific. But his and Messick's definition of validity does not mention measurement. Scientific metrology is only about measurement. Kane has written about how lawsuits will force reality where there is inadequate evidence to support claims being made. He is being true to the pragmatic applied and legal position he has taken on what is important in validation and often gives good scientific advice at other times. While not defining measurement or the instrument as involved in validity, Kane speaks to the value, but higher cost, of validation of constructs and their theory. What might be most puzzling in his approach, when identified with applied and legal metrology, is that these are precisely the domains for which instrument developers will be held accountable. Interpretations and actions are fully in the information world. Similarly psychometric experts Wilson and Maul, as advocates of their VIM-centered definition of measurement, give excellent advice for applied concerns in their other writings.

Scientific metrology includes these higher-cost studies but goes farther, leading to acceptance of standard units for all participating nations. Educational and psychological measurements do not practice this aspect of scientific metrology currently. *The Open Skills Network* mentioned at the end of Chapter 8 is a way of establishing commonly accepted definitions of skills in the college-level development of rich skill descriptions. But these RSDs are not yet measurement scales. There is a long distance to be traveled before any scales that measure such skills will have a scientific unit defined so that any new instrument for assessing mastery of that skill could be traceable to the unit's scientific definition. Meanwhile, the RSD provides a common language so important to employers, future employees, and educators who are trying to assure that students possess the competencies employers need and seek.

Despite the calm assurance of spokespersons from these extremes about the strength of their positions, a vibrant validity theory must span all the way from the scientific end to the legal end of the metrology spectrum. That is another reason that this book supports the Mislevy (2018, p. 202) definition that *does span* total metrology: "A test is valid for a given interpretation or use to the degree to which empirical evidence and theoretical rationales support reasoning as if *it measures what it is purported to measure."* This sociocognitive foundations book by Mislevy (2018) provides an entire chapter 8 on validity, and surrounding parts of this chapter on reliability, fairness (part of validity), comparability, and the meanings of terms in various quantitative measurement models. In so doing, he gives the nuanced view from the perspective of a sociocognitive constructivist-realist position. Like his broadened but nuanced definition, his Chapter 8 and his entire book broaden the concept of measurement and its role with people from varying cultural, linguistic, and educational backgrounds also, at situated moments in time when they are assessed by a measurement instrument for some set of purposes.

9.4.3 The Western Governors University Experiment and the Ostensive Demonstration of Skills

In Chapter 8, ostensive definitions and demonstrations were defined within the *Educational & Psychological Measurement* domain to be test-taker-friendly in multiple ways. First, they arise from domain analyses that locate the primary values and consequences of relevance to the test taker's future aspirations. Second, the test taker is permitted to repeat the same ostensive demonstration more than once, receiving coaching and guidance after each assessed performance. By coaching and guidance, we mean interpretations and advice for action. These are the applied and legal aspects of validation. This feedback must occur shortly after each competency testing occasion to enable optimal learner progress. We showed how implementing this very student-friendly use

of measurement requires a profound revision of how education is organized and operates. We discussed the example of WGU which has created a reorganization of teaching roles and student roles. We showed that its growth was beyond remarkable, from one graduate in 2000 to 290,000 in the summer of 2022. We promised some data to point the eyes of new psychometricians toward where the market for assessment development and management is going in *Educational Measurement*. Competency-based learning using ostensive definition and demonstration—by whatever name—will survive and thrive. We conclude with some evaluation results and statistics for the argument that you would be on strong ground to follow the way economics and user demand will be taking *Educational Measurement and Psychometrics* in the future.

WGU traditions are to evaluate at least annually the competency of the organization in attaining its goals and purpose, following the example of what its students are to do. The exams for the University as a whole are built up from the accomplishments compared to goals within each of its four colleges, Teacher Education, Business, Information Technology, and Health Professions. The annual reports have been the time to share the results of these college and university-wide evaluations. Special occasions such as the 20th anniversary and the 25th anniversary merit special reports. Table 1 is a high-level summary of the growth in cumulative number of undergraduate and master's degrees through 31 March 2023.

An update on the most recent numbers is found at *WGU's Story*, available at: https://www.wgu.edu/about/story/measuring-impact.html.

This web page provides information up-to-date as of 31 March 2023. The rapid growth continues.

- Diverse Student Body—**143,959** in all 50 states and at Military Bases around the world.
- Alumni—**317,107** accomplished professionals in all 50 states and around the world.

For WGU's 25th anniversary, the Gallup Organization was asked to develop an Impact Report by asking over 1400 WGU graduates a series of questions. They were asked about their WGU experience and what it has meant to them since graduation. We briefly summarize this report here.

At mid-year of its 25th anniversary, WGU exceeded 290,000 graduates with most of these degrees having gone to nontraditional students several years removed from high school, financially on their own, and married. Compared to college students nationally over this period, these groups are strongly under-represented in higher education. WGU graduates are more likely to work while enrolled, and two-thirds are from underserved populations including graduates of color, or those from first-generation families, low-income backgrounds, or rural areas. The mix of majors at WGU is oriented toward skills that lead to improved career outcomes, including higher average salaries and greater employer satisfaction. WGU participates in the open skills network and communicates with employers in this and other ways to determine what rich skills descriptions are going to be needed when their students graduate, and what skills/competencies are increasing and decreasing in demand among employers.

On average, WGU students paid less than two-thirds of the tuition of the average American undergraduate. This lowers costs and maximizes convenience, allowing parents, workers, and

Table 1 Western Governors University growth in cumulative graduates and annual enrollment. Graduates include both undergraduate and master's degrees.

Year	2000	2005	2014	2015	2016	2017	2018	2019	2020	2021	2022
Graduate	1	500	41,499	56,592	75,010	98,765	128,894	167,173	214,194	261,569	290,000
Enrolled	300	5000	50,000	60,000	76,743				129,169	128,394	

others an opportunity to balance personal responsibilities with academic work. Along with the relatively low cost of tuition and WGU's flat-rate-per-term tuition model, WGU students progress more rapidly through the process of degree attainment, further mitigating costs and debt for WGU students. The Gallup Impact Report estimated that, in 2021, WGU alumni contributed $8 billion more to the US economy than they would have if they never attended college. The report also found that WGU alumni are more likely to be thriving in their current lives and more engaged in their jobs than bachelor's degree holders or adults without a postsecondary degree nationally.

9.5 Summary

This documents the sensitivity of WGU both to the situation and needs of their student population, and their sensitivity to the market for the competencies employers need. It also documents that employers are very satisfied with the competency and work ethic of their WGU employees. The claimed competencies, assured by strong and valid assessments, are standing up to or exceeding the employers' expectations. These assessments rely on ostensive demonstrations of ability to perform in ways that transfer to work performance.

The "obsessive" focus on student success is not wavering in the WGU culture. Thus, the resultant remarkable growth and student satisfaction is not slowing down nor is it likely to do so. The inference that this is wholly due to ostensive demonstrations of competency is of course not being claimed in this book. Another factor is the fact that the WGU student population has largely been underserved. But they are more mature and focused. However, all these variables relate to the remarkable results, it raises a challenge for traditional ways of organizing post-secondary education and assessing outcomes. This is placing a challenge on traditional colleges, and pre-colleges, that cannot deliver the success ratio, much lower cost, flexibility, or speed to a degree that WGU has demonstrated it can deliver.

Our point is that external validity, the performance on the job, can be assured by using many more ostensive performance demonstrations along the way and reorganizing to replace nonstudent-centered old ways with new arrangements. We are inviting those who graduate in psychometrics to think hard about the uses of testing that might be losing market share as other uses of assessment are threatening to make them irrelevant.

We have also put assessment in the framework of international metrology, spanning scientific, applied, and legal aspects. Doing so has clarified the source of unnecessary conflict in the theory of validity of assessment, and can shed light in many other areas. It has opened up new possibilities that someday units and scales will be standardized for valuable attributes of human beings. The focus on international metrology will have to prove its merit in the world's existing ostensive ways to standardize high-value performances. These methods have been going on for centuries in the Olympics performance domains, and in professional sports, performing arts, and more recently, creative and complex new games that may predict competency in useful occupations or work. These areas have been setting up their own metrics for years. The traditions of international scientific metrology ought at least to shine a light on these performance domains that assess and coach for high levels of performance. Meantime, science is coming in at least in ways to maximize team success through various forms of domain analysis and domain modeling unconnected to the ECD methods we have encouraged in educational and psychological assessment development.

This current writing team has worked on creating this book over the past decade, and Petr for a decade or more before that, to bring about this unlikely and unusual book. We have tried to make it learner-friendly in as many ways as possible, yet challenging and referenceable for years. We have tried to connect it with the current issues in measurement, especially international metrology, and the topics of *Psychometrics*, *Test Theory*, and *the Latent Factor Model*.

References

Borsboom, D. (2018). *The philosophy of psychometrics*. An invited presentation to the Psychometric Society https://vimeo.com/256145513.

Borsboom, D., Mellenbergh, G.J., and Van Heerden, J. (2004). The concept of validity. *Psychological Review 111*: 1061–1071.

Choi, Y.J. and Asilkalkan, A. (2019). R packages for item response theory analysis: descriptions and features. *Measurement: Interdisciplinary Research and Perspectives 17*: 168–175.

Downing, S.M. and Haladyna, T.M. (ed.) (2006). *Handbook of test development*, firste. Routledge.

JCGM (2012). *International Vocabulary of Metrology—Basic and general concepts and associated terms (VIM)*. (3rd ed. 2008). Joint Committee for Guides in Metrology (www.bipm.org/en/publications/guides/vim.html).

Mari, L., Wilson, M., and Maul, A. (2021). *Measurement across the sciences. Springer series in measurement science and technology*. Springer.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

Mislevy, R.J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Mislevy, R. and Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology (PADI Technical Report 9)*. Menlo Park, CA: SRI International.

Simon, H.A. (1996). *The sciences of the artificial*. The Massachusetts Institute of Technology.

Computational Guide 1

Classical Test Theory

A Model-Based Approach

CHAPTER MENU

1 Classical Test Theory and the True Score Model, 360
  1.1 Reliability with Multiple Test Components, 360
  1.2 True Score Models for Multiple Items, 361
  1.3 Latent Variable Analysis for True Score Models, 361
  1.4 True Score Models as Factor Models, 363
  1.5 Latent Variable Modeling of a Composite Score, 365
  1.6 Inter-Item and Item-Total Correlations, 366
2 Estimating True Score Models as Latent Variable Models, 368
  2.1 The Congeneric Model, 368
  2.2 Tau-Equivalent and Parallel Models, 370
  2.3 A Strictly Congeneric Model, 371
  2.4 Item Reliability, 371
  2.5 The Congeneric Model with Equally Reliable Indicators, 372
3 Composite Reliability, 374
  3.1 Coefficient Alpha, 375
  3.2 Bootstrapping Coefficient Alpha, 376
  3.3 The Reliability Index for Congeneric Data, 376
  3.4 Alternate Estimation of Coefficient Alpha, 378
  3.5 Coefficient Omega, 379
  3.6 Bootstrapping Omega, 380
  3.7 Jointly Estimating and Comparing Alpha and Omega, 380
  3.8 Omega If an Item Is Deleted, 382
  3.9 Coefficient Rho, 383
  3.10 Maximal Reliability and Optimally Weighted Indicators, 384
4 Standardized Reliability, 385
  4.1 Standardized Alpha, 386
  4.2 Standardized Omega, 387
  4.3 Standardized Rho, 389
  4.4 Standardized Maximal Reliability (Coefficient *H*), 390
5 Summary, 391
References, 397

This is the first of three computational guides addressing practical problems and common tasks in psychometric analysis. This first guide covers Classical Test Theory while the second and third cover Item Response Theory and Factor Analysis, respectively. These guides are intended to give the reader an opportunity to apply basic concepts from the book to empirical situations using statistical modeling software. They provide how-to-do-it instructions for performing a variety of psychometric and statistical tasks and activities. This first guide focuses on how to estimate various

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

statistical models and accomplish certain other activities related to basic Classical Test Theory (CTT). These are illustrated primarily using Mplus, although a similar code for Stata is also provided for some of them. The examples in the guide can be completed using the free demonstration version of Mplus available from the Mplus website http://statmodel.com/demo.shtml.

After introducing CTT and associated true score models, latent variable common factor modeling is used to estimate parallel, tau-equivalent, and congeneric test models. Reliability assessment for composite scores is outlined using a variety of reliability coefficients, including alpha, omega, rho, and maximal reliability. We also address bootstrapping of reliability coefficients and reliability estimation with standardized indicators and describe the assessment of reliability if items were omitted.

1 Classical Test Theory and the True Score Model

In the traditional true score model, the observed score Y_i for individual i is treated as the sum of a true score T_i and a measurement error E_i representing the departure of the observed score from the true score.

$$Y_i = T_i + E_i$$

It is usually assumed that the true scores are uncorrelated with the measurement errors, and that the measurement errors are uncorrelated with each other. Under these conditions, the variance of the observed scores is:

$$Var(Y_i) = Var(T_i) + Var(E_i)$$

Because the measurement errors are also assumed to be distributed with a mean of zero and a variance of $Var(E_i)$, the expected value of the true scores is equal to the expected value of the observed scores.

$$E(T_i) = E(Y_i)$$

Reliability is defined as the ratio of the variance of the true scores to the variance of the observed scores.

$$R = \frac{Var(T_i)}{Var(Y_i)} = \frac{Var(T_i)}{Var(T_i) + Var(E_i)}$$

In practice, a single observed score cannot be uniquely partitioned into true score and error components and their associated variances, but if multiple measurements are available, an estimate of the true score variance can be obtained. For example, if there are two measurements

$$Y_{i1} = T_i + E_{i1}$$
$$Y_{i2} = T_i + E_{i2}$$

and the two error variances are further assumed to be equal, $Var(E_{i1}) = Var(E_{i2}) = Var(E_i)$, the measurements are said to be parallel, and the covariance between the two observed measurements is equal to the true score variance.

$$Cov(Y_{i1}, Y_{i2}) = Cov(T_i + E_{i1}, T_i + E_{i2}) = Var(T_i)$$

1.1 Reliability with Multiple Test Components

The corresponding correlation between the two component scores is then equal to the measure's reliability. When the two component scores are obtained from two different raters, the reliability is referred to as inter-rater reliability. When the two parallel component scores are obtained on two

different occasions, it is called test-retest reliability. When it is obtained from two subtests, it is called split-half reliability.

$$Corr(Y_{i1}, Y_{i2}) = \frac{Cov(Y_{i1}, Y_{i2})}{\sqrt{var(Y_{i1})}\sqrt{var(Y_{i2})}} = \frac{Var(T_i)}{Var(T_i) + Var(E_i)} = R$$

Split-half reliability can be considered a rudimentary form of so-called internal consistency reliability. Oftentimes, the interest is not in the estimated reliability of a given half-test, but in the estimated reliability of the composite score obtained as the sum of the two half-tests. With $J = 2$ subtests, the reliability of the composite score can be obtained from the Spearman–Brown formula:

$$R_{SB} = \frac{Jr}{(J-1)r + 1} = \frac{2r}{r+1}$$

For example, if the split-half correlation r were 0.80, the estimated Spearman–Brown reliability for the combination of the two half-tests would be $2(0.80)/(0.80 + 1) = 0.89$.

1.2 True Score Models for Multiple Items

Where there are multiple test replications or components that share a common true score, the model for the score of individual i on component j can be expressed as:

$$Y_{ij} = T_i + E_{ij}$$

With the additional restriction of homogeneous residual variances, $Var(E_{ij}) = \sigma^2_\varepsilon$, this is described as the *(strictly) parallel* model. The true score model with a common residual variance can be generalized to permit subtest j to have a distinct offset from the true score by an additive constant a_j.

$$Y_{ij} = a_j + T_i + E_{ij}, \;\; Var(E_{ij}) = \sigma^2_\varepsilon$$

This is described as the *essentially parallel* model. If the error variances can also vary from item to item, $Var(E_{ij}) = \sigma^2_{\varepsilon_j}$, the model is considered to be *(essentially) tau-equivalent*, indicating that the test components have equivalent true scores.

$$Y_{ij} = a_j + T_i + E_{ij}, \quad Var(E_{ij}) = \sigma^2_{\varepsilon_j}$$

Finally, the model can be further generalized to permit each subtest j to be related to the true score through a subtest-specific multiplicative constant b_j. This is described as the *congeneric* model or the true score model for congeneric items.

$$Y_{ij} = a_j + b_j T_i + E_{ij}, \quad Var(E_{ij}) = \sigma^2_{\varepsilon_j}$$

The parallel and tau-equivalent models assume the test components share a common true score. The test components in the *strictly parallel* model are further assumed to have the same mean and (residual) variance. The *essentially parallel* model assumes that the test components have a common residual variance but allows the means to differ. In the *essentially tau-equivalent* model, subtests can have differing means and residual variances but still possess a common true score contribution. In the *congeneric* model, the true score contribution can vary across test components, as can the means and residual variances.

1.3 Latent Variable Analysis for True Score Models

This extended true score model is identical in form to a single-factor latent variable model.

$$Y_{ij} = \nu_j + \lambda_j \eta_i + \varepsilon_{ij}$$

If there are four subtests, for example, this model can be expressed in matrix form for a given individual i as:

$$\mathbf{Y} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \end{bmatrix} + \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} [\eta_i] + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix}$$

Here, y_{ij} is an observed score on subtest j for individual i, ν_j is the intercept for subtest j, λ_j is the factor loading for subtest j on the latent variable η_i, and ε_{ij} is the subtest j residual for individual i.

In addition to the subtest-specific factor loadings, λ_j, and intercepts, ν_j, the other parameters of the model are subtest-specific residual variances, $Var(\varepsilon_{ij}) = \theta_{jj}$, the variance of the latent variable, $Var(\eta_i) = \psi$, and the mean or expected value of the latent variable, $E(\eta_i) = \alpha$. The expectations of the errors are fixed at zero, $E(\varepsilon_{ij}) = 0$, and their variances are estimated as model parameters but all covariances among the errors are fixed to zero.

$$Cov(\varepsilon_j) = Cov \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix} = \begin{bmatrix} \theta_{11} & 0 & 0 & 0 \\ 0 & \theta_{22} & 0 & 0 \\ 0 & 0 & \theta_{33} & 0 \\ 0 & 0 & 0 & \theta_{44} \end{bmatrix}.$$

The model-implied covariance matrix is then defined as:

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$$

$$\begin{bmatrix} \hat{\sigma}_{11}^2 & \hat{\sigma}_{12} & \hat{\sigma}_{13} & \hat{\sigma}_{14} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22}^2 & \hat{\sigma}_{23} & \hat{\sigma}_{24} \\ \hat{\sigma}_{31} & \hat{\sigma}_{32} & \hat{\sigma}_{33}^2 & \hat{\sigma}_{34} \\ \hat{\sigma}_{41} & \hat{\sigma}_{42} & \hat{\sigma}_{43} & \hat{\sigma}_{44}^2 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} [\psi] [\lambda_1 \quad \lambda_2 \quad \lambda_3 \quad \lambda_4] + \begin{bmatrix} \theta_{11} & 0 & 0 & 0 \\ 0 & \theta_{22} & 0 & 0 \\ 0 & 0 & \theta_{33} & 0 \\ 0 & 0 & 0 & \theta_{44} \end{bmatrix}$$

The model-implied covariance matrix $\hat{\boldsymbol{\Sigma}}$ is symmetric about its principal diagonal. For example, $\hat{\sigma}_{21} = \hat{\sigma}_{12}$. The expectations of the observed variables are $E(Y_{ij}) = \hat{\mu}_j = \nu_j + \lambda_j\alpha$, and the model-implied mean vector for the indicators is thus:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\alpha}$$

$$\begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \\ \hat{\mu}_4 \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \end{bmatrix} + \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} [\alpha]$$

To identify the factor model and scale the latent variable η, restrictions are often imposed on certain model parameters. One common method, as shown in Figure 1.1, is to fix the variance of the latent variable to 1, $\psi = 1$, and fix the latent variable mean to zero, $\alpha = 0$. When the latent variable's variance is fixed to 1, the model-implied covariance structure can be given as follows.

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} [1] [\lambda_1 \quad \lambda_2 \quad \lambda_3 \quad \lambda_4] + \begin{bmatrix} \theta_{11} & 0 & 0 & 0 \\ 0 & \theta_{22} & 0 & 0 \\ 0 & 0 & \theta_{33} & 0 \\ 0 & 0 & 0 & \theta_{44} \end{bmatrix} = \begin{bmatrix} \lambda_1^2 + \theta_{11} & \lambda_1\lambda_2 & \lambda_1\lambda_3 & \lambda_1\lambda_4 \\ \lambda_2\lambda_1 & \lambda_2^2 + \theta_{22} & \lambda_2\lambda_3 & \lambda_2\lambda_4 \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & \lambda_3^2 + \theta_{33} & \lambda_3\lambda_4 \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & \lambda_4^2 + \theta_{44} \end{bmatrix}$$

When the latent mean is fixed to zero, the model-implied means are given by the indicator intercepts.

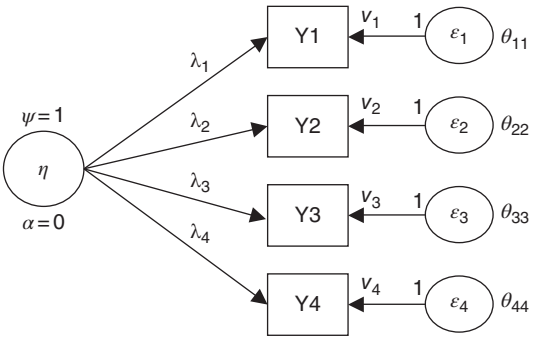


Figure 1.1 Path diagram for estimating true score models with latent variable modeling. As in similar path diagrams, circles are used to represent latent variables or residuals, rectangles are used for observed variables, and straight single-headed arrows are used to represent regression relationships (including factor loadings).

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \\ \hat{\mu}_4 \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \end{bmatrix} + \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} [0] = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \end{bmatrix}$$

1.4 True Score Models as Factor Models

Beyond the constraints necessary to identify the model and scale the latent variable, various true score models can be estimated by imposing certain additional equality constraints on the loadings, intercepts, and residual variances (see Table 1.1). The most highly constrained model is the *strictly parallel* model with equal loadings, intercepts, and residual variances. The *essentially parallel* model equates the loadings and residual variances but permits indicator-specific intercepts. The *strictly tau-equivalent* model equates the loadings and intercepts but permits indicator-specific residual variances. The *essentially tau-equivalent* model equates the loadings but permits indicator-specific intercepts and residual variances. The *strictly congeneric* model freely estimates the latent variable mean while fixing the indicator intercepts to zero, and estimating indicator-specific loadings and residual variances. The least constrained model is the *(essentially) congeneric* model that freely estimates indicator-specific loadings, intercepts, and residual variances. For J indicators, the number of parameters for these models ranges from 3 for the (most parsimonious) strictly parallel model to $3J$ for the (most general) congeneric model.

Certain of these models are nested and can be compared using standard model comparison procedures for nested models such as a Chi-square difference test (see Figure 1.2 and Table 1.2). For example, comparing the congeneric model with the essentially tau-equivalent model tests the equality of the factor loadings. Comparing the essentially tau-equivalent model with the

Table 1.1 Estimating various true score models for J indicators when identifying and scaling the latent variable by constraining the factor variance $\psi = 1$.

Factor loadings	Factor mean	Indicator intercepts	Residual variances	Number of parameters	True score model
$\lambda_j = \lambda$	$\alpha = 0$	$\nu_j = \nu$	$\theta_j = \theta$	3	(Strictly) parallel
$\lambda_j = \lambda$	$\alpha = 0$	ν_j	$\theta_j = \theta$	$J + 2$	Essentially parallel
$\lambda_j = \lambda$	$\alpha = 0$	$\nu_j = \nu$	θ_j	$J + 2$	(Strictly) tau-equivalent
$\lambda_j = \lambda$	$\alpha = 0$	ν_j	θ_j	$2J + 1$	Essentially tau-equivalent
λ_j	α	$\nu_j = 0$	θ_j	$2J + 1$	(Strictly) congeneric
λ_j	$\alpha = 0$	ν_j	θ_j	$3J$	Essentially congeneric

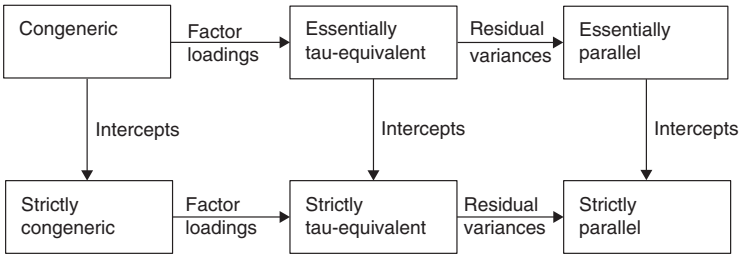


Figure 1.2 True score model nesting and model comparisons.

Table 1.2 One-step testing model parameter equality.

Test equality of	More general model	More restricted model	Conditional upon equal
Factor loadings	Essentially congeneric	Essentially tau-equivalent	
Intercepts	Essentially congeneric	Strictly congeneric	
Factor loadings	Strictly congeneric	Strictly tau-equivalent	Intercepts
Intercepts	Essentially tau-equivalent	Strictly tau-equivalent	Factor loadings
Residual variances	Essentially tau-equivalent	Essentially parallel	Factor loadings
Residual variances	Strictly tau-equivalent	Strictly parallel	Loadings and intercepts
Intercepts	Essentially parallel	Strictly parallel	Loadings and variances

essentially parallel model tests the equality of the residual variances, as does comparing the strictly tau-equivalent model with the strictly parallel model. Comparing the essentially tau-equivalent model with the strictly tau-equivalent model tests the equality of the intercepts, as does a comparison of the essentially parallel model with the strictly parallel model.

If the simpler tau-equivalent or parallel models fit the data as well as the congeneric model, model parsimony is improved, and model stability may be enhanced. Fitting the various true score models can also be useful and informative with respect to reliability assessment and scale quality evaluation. For example, at least essential tau-equivalence is necessary in order for Cronbach's Coefficient Alpha to constitute a genuine lower bound to the population reliability coefficient, and a good fit of the essentially parallel model is required in order to appropriately apply the Spearman–Brown prophecy formula to forecast the reliability of a shortened or lengthened test.

Congeneric models will generally fit measurement data better than the more restrictive tau-equivalent or parallel models. On the other hand, when tau-equivalent or parallel models fit well, they provide potentially useful model simplification that may replicate better in other samples. Congeneric models are especially useful where the data have not been obtained using the same response format for all indicators, or where the data have not been previously equated to a common scale. With occasional exceptions, researchers and methodologists are commonly referring to "essentially" parallel, tau-equivalent, and congeneric models, respectively, when they talk about "the" parallel, tau-equivalent, or congeneric model.

We have included the "strictly" parallel, tau-equivalent, and congeneric models to provide a more complete treatment (see also O'Grady & Medoff, 1991; Millsap & Everson, 1991), and we consider potentially useful applications of these models. The distinction between strictly and essentially parallel and tau-equivalent true score models is generally well recognized in CTT. Strictly or randomly parallel measures share a common mean and are fully exchangeable while essentially parallel measures can have differing means or offsets. Similarly, strictly tau-equivalent measures are expected to

possess a common mean, though possibly differing in their residual variances, while essentially tau-equivalent measures can differ in their means as well as the residual variances.

Our inclusion of a strictly congeneric model is much less common in CTT treatments of measurement modeling. This model considers a possibly nonarbitrary relation between an estimated latent (true score) mean and the expected or model-implied intercepts, namely that once the influence of the latent mean is taken into account via the factor loadings, further intercept deviations will not be expected to differ from zero. In some ways, this is analogous to the idea of local independence where once we have controlled for the influence of the underlying latent variable, we no longer expect correlations among the measure residuals. As with the strictly parallel and strictly tau-equivalent models, the mean structure of the data will be represented with a single parameter, rather than having a saturated mean structure. Estimating the strictly congeneric model and comparing it to the usual congeneric model has potential applications in linking, equating, or calibrating measures as well as theoretical measurement model fit assessment. The fit of the strictly congeneric model will always be less than or equal to the fit of the familiar (essentially) congeneric model when judged by a model fit Chi-square test.

Before demonstrating how to estimate these and some other measurement models using the confirmatory factor analysis and structural equation modeling (SEM) software Mplus, we will introduce the use of Mplus to estimate basic inter-item and item-total correlations. Although these correlations are easily obtained from almost any statistical software package, or even a spreadsheet such as Excel, we want to do this within the framework of statistical modeling software, particularly SEM software. Although we will use Mplus, and will exploit special convenience features of Mplus, many of these methods are also readily implemented in other SEM software such as Stata's sem and gsem, SAS's Proc Calis, and Amos from SPSS, as well as R packages like lavaan or openmx. Among the reasons for taking this approach is the ability to incorporate modern missing data handling strategies such as full information maximum likelihood (FIML) instead of common default methods like listwise deletion.

1.5 Latent Variable Modeling of a Composite Score

For certain purposes, it can be convenient to incorporate an equally weighted or differentially weighted composite of the items or indicators as an ancillary feature of traditional latent variable analysis of true score models. This can be accomplished in many SEM software programs by creating a so-called "phantom" latent variable with no indicators and a zero intercept and residual variance (see Figure 1.3). The coefficients for regressing the phantom composite latent variable on the indicators are fixed at 1 to create a unit-weighted composite as the sum of the indicators. The latent variable η_0 does not affect model fit or the estimation of model parameters, but most SEM software allows users to request the implied means and variances of the composite, along with covariances and correlations of the composite with the indicators.

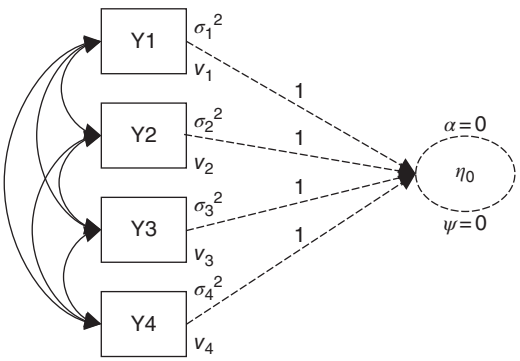


Figure 1.3 Latent variable modeling of a composite score.

1.6 Inter-Item and Item-Total Correlations

We illustrate the estimation of these models in Mplus with data from Peters (2014)*. Jointly estimating the associations among test items or indicators and their total or sum score is part of what is sometimes called *item analysis*. With only the individual indicators as input, the correlations among the items and the correlations of the items with the composite (scale) score can be obtained using the following Mplus syntax [ItemTotalUncorrected.inp].

```
TITLE: Inter-item and (uncorrected) item-total correlations
 Also scale and item variances and means
DATA: FILE IS; selectedExampleData.dat
VARIABLE: NAMES ARE v01 v02 v03 v04 v05 v11 v12 v13 v14 v15;
 USEVARIABLES ARE v01-v05;   ! Select variables for analysis
MODEL: v01-v05 WITH v01-v05; ! Covariances
 s BY;                       ! Composite variable
 s@0;                        ! Set the variance of the composite to 0
 s ON v01-v05@1;             ! Create equally weighted composite
OUTPUT: STDYX TECH4;         ! Uncorrected item-total correlations
```

Because this is the first time we have encountered Mplus modeling syntax, we should introduce the commands shown here, noting that some are fairly self-explanatory. The optional TITLE: command gives one or more lines of descriptive text about the model, the DATA: command specifies the name of the data file, the VARIABLE: command specifies the variable names, and the USEVARIABLES subcommand selects a subset of the variables for analysis. Mplus commands and subcommands are terminated with a semicolon (;) and anything on a line following an exclamation point (!) is treated simply as a comment and ignored during syntax execution.

The MODEL command first specifies that the five initial indicator variables, v01 to v05, are to be correlated WITH one another. The next three lines define the phantom composite variable **s** (representing the overall sum of the items), fix its variance to zero (its mean is fixed to 0 by default), and regress it ON the indicators with regression coefficients fixed to one using the @ sign. The STDYX specification on the OUTPUT: command requests standardized results, and the TECH4 specification requests estimates of the means and variances of the model implied equally weighted (phantom variable) composite scale score (not shown), and the correlations of the scale score with the indicators. This is a saturated model with all possible indicator covariances. The raw or uncorrected item-total correlations are underlined and bolded in selected output from the TECH4 section of the Mplus output shown as follows.

---

* To gain skill using the Mplus code segments given here, download the free demonstration version of Mplus available from the Mplus website: http://statmodel.com/demo.shtml

To obtain the dataset provided for the MPlus demonstration code segments, copy this https: location which is specified in the last 21 characters of the Peters (2014) reference, paste them in a web browser and go to this location: https://osf.io/tnrxv/

This will take you to the Open Science Foundation (OSF) website where the Peters (2014) paper is located. On the selection ribbon located above the title of the Peters paper select "files." From the list of possible downloads, click on the sixth one, the one entitled "selectedExampleData.csv." Save this *.csv* file as a *.txt* file. Delete the first line, which consists of labels for the ten columns of "items." (Mplus files are not to have labels. They consist of only numerical values.)

Be sure to place this dataset file in the same folder in which the Mplus code files copied from this computational guide are placed. If you do so, whenever you run any one of these Mplus code files it will read the "Data:" line in the Mplus header input that specifies this particular data file and will execute the instructions of the code using this dataset.

```
TECHNICAL 4 OUTPUT
...        ESTIMATED CORRELATION MATRIX FOR THE LATENT VARIABLES
            S              V01            V02            V03            V04
______      ________       ________       ________       ________       ______
 S           1.000
 V01         0.665          1.000
 V02         0.830          0.488          1.000
 V03         0.673          0.509          0.432          1.000
 V04         0.774          0.474          0.589          0.450          1.000
 V05         0.809          0.435          0.465          0.494          0.470
```

An important advantage of this latent composite approach is that it can appropriately handle cases with missing values on the indicators through applying FIML methods when estimating the item-total correlations. However, this "uncorrected" item-total correlation can potentially be inflated by the part-whole relationship between the item response and the ordinary total scale score. In this case, the correlation between the response to a given item and the sum of the remaining items, sometimes called the item-rest correlation or the "corrected" item-total correlation, may be desired instead. The following modifications to the MODEL command in the previous Mplus Syntax can be used to apply the FIML latent composite method to the estimation of corrected item-total correlations. In each case, separate "total" scores (s1–s5) are created for each item as the sum of the remaining items [ItemTotalCorrected.inp].

```
! Corrected item-total correlations
MODEL: v01-v05 WITH v01-v05;       ! Covariances
 s1 BY v01; s2 BY v02; s3 BY v03; ! Corrected latent composite
 s4 BY v04; s5 BY v05;            !  variables
 s1-s5 ON v01-v05@1;              ! Composite scores for each item
 s1-s5 PON v01-v05@0;             ! Correct the composites
 s1-s5@0;                         ! Set composite variances to 0
 s1-s5 WITH s1-s5@0;              ! Set composite covariances to 0
OUTPUT: STDYX TECH4; ! Tech4 gives corrected item-total correlations
```

The corrected item-total correlations are located on the diagonal of the matrix containing the correlations between the items and the corrected composites, and are underlined and bolded in the following section of the Mplus output.

```
TECHNICAL 4 OUTPUT
    ESTIMATES DERIVED FROM THE MODEL
          ESTIMATED CORRELATION MATRIX FOR THE LATENT VARIABLES
            S1            S2            S3            S4            S5
            ________      ________      ________      ________      ________
 V01        0.591         0.668         0.650         0.665         0.685
 V02        0.835         0.624         0.844         0.832         0.899
 V03        0.662         0.710         0.582         0.682         0.664
 V04        0.776         0.768         0.780         0.644         0.818
 V05        0.818         0.882         0.811         0.840         0.577
```

Table 1.3 Inter-item and uncorrected and corrected item-total correlations for Peters' data ($N = 250$).

	Total scale scores		Scale component (item) scores				
Items	**Uncorrected**	**Corrected**	**V01**	**V02**	**V03**	**V04**	**V05**
V01	**0.665**	**0.591**	1.000				
V02	**0.830**	**0.624**	0.488	1.000			
V03	**0.673**	**0.582**	0.509	0.432	1.000		
V04	**0.774**	**0.644**	0.474	0.589	0.450	1.000	
V05	**0.809**	**0.577**	0.435	0.465	0.494	0.470	1.000

Table 1.3 presents the corrected and uncorrected item-total correlations with the inter-item correlations in a single display.

In each case here, the corrected item-total correlations are lower than the uncorrected ones.

2 Estimating True Score Models as Latent Variable Models

Following common practice, we have described the partitioning of the total test variance into true score variance and error variance in true score models and the similar partitioning of the total observed variance into common variance and error variance in common factor models. For an additional perspective on further partitioning of what we have thus designated as error variance, see Section 64 in Chapter 6 of the main text. Graham (2006) and others have outlined congeneric, tau-equivalent, and other test score models that we introduce and demonstrate in the following sections.

2.1 The Congeneric Model

The general congeneric model can be estimated using the following Mplus syntax [EssentiallyCongeneric.inp]:

```
TITLE: (Essentially)Congeneric Model
DATA: FILE IS selectedExampleData.dat;
VARIABLE: NAMES ARE v01 v02 v03 v04 v05 v11 v12 v13 v14 v15;
 USEVARIABLES ARE v01-v05; ! Select variables for analysis
MODEL: f BY v01-v05*;      ! Indicator-specific factor loadings
 f@1;                      ! Fix the factor variance at 1
OUTPUT: STDYX;             ! Standardized results
```

Here, the MODEL: command uses an asterisk (*) to specify that the latent variable f is measured BY five indicators, whose factor loadings are all freely estimated. The next line fixes the variance of the latent variable (the common factor) to one. By default, Mplus also freely estimates item-specific residual variances and indicator intercepts. The STDYX option on the OUTPUT: command requests the completely standardized results, which standardize the observed and latent variables by rescaling their variances to 1.0. This model fits the data quite well with both the CFI and TLI

above 0.95 and the SRMR below 0.05. Although the RMSEA is above 0.05, the 90% confidence interval includes 0.05, and the associated test of approximate fit is not statistically significant.

```
Chi-Square Test of Model Fit
          Value                             13.260
          Degrees of Freedom                    5
          P-Value                           0.0211
RMSEA (Root Mean Square Error Of Approximation)
          Estimate                           0.081
          90 Percent C.I.                    0.029  0.136
          Probability RMSEA
CFI/TLI
          CFI                                0.979
          TLI                                0.958
Chi-Square Test of Model Fit for the Baseline Model
          Value                            402.737
          Degrees of Freedom                   10
          P-Value                           0.0000
SRMR (Standardized Root Mean Square Residual)
          Value                              0.027
```

The unstandardized and standardized factor loadings and the item reliabilities (R-squared values) from this specification of the congeneric model are given in the following selection from the Mplus output.

```
                    Estimate       S.E.  Est./S.E.    P-Value
 F        BY
    V01                1.982      0.178     11.133      0.000
    V02                6.914      0.572     12.093      0.000
    V03                2.341      0.218     10.719      0.000
    V04                4.369      0.359     12.174      0.000
    V05                6.423      0.598     10.737      0.000
...
STDYX Standardization
 F        BY
    V01                0.681      0.043     15.962      0.000
    V02                0.727      0.040     18.398      0.000
    V03                0.664      0.044     14.959      0.000
    V04                0.731      0.039     18.619      0.000
    V05                0.662      0.044     15.031      0.000
R-SQUARE
    V01                0.464      0.058      7.981      0.000
    V02                0.529      0.058      9.199      0.000
    V03                0.441      0.059      7.480      0.000
    V04                0.534      0.057      9.309      0.000
    V05                0.438      0.058      7.516      0.000
```

2.2 Tau-Equivalent and Parallel Models

It is useful to remember that by default Mplus estimates item-specific intercepts and residual variances, and item-specific factor loadings are obtained by using the asterisk (*) following the first indicator name, v01* v02 v03 v04 v05, or following the series when the indicators are specified as a series, v01–v05*. The essentially tau-equivalent model requires equating the factor loadings while allowing Mplus' default estimation of indicator-specific intercepts and residual variances. The essentially tau-equivalent model is obtained by modifying the congeneric model's Mplus MODEL: command as follows [EssentiallyTauEquivalent.inp]:

```
! Essentially Tau Equivalent model
MODEL: f BY v01-v05* (b);  ! Common factor loadings
 f@1;                      ! Fix the factor variance at 1
```

The (b) in parentheses following the factor specification indicates that the five-factor loadings are constrained to be equal. The strictly tau-equivalent model requires equating the factor loadings and the indicator intercepts but still (by default) allowing item-specific indicator residual variances. Square brackets, [], are used to identify means and intercepts in Mplus. The (a) following the bracketed specification of the intercepts equates the five indicator intercepts to obtain the strictly tau-equivalent model [StrictlyTauEquivalent.inp].

```
! Strictly Tau Equivalent model
MODEL: f by v01-v05* (b);  ! Common factor loading
 f@1;                      ! Fix the factor variance at 1
 [v01-v05] (a);            ! Common intercept
```

The essentially parallel model requires equating the factor loadings and indicator residual variances while allowing indicator-specific intercepts. In Mplus, variances and residual variances are designated by listing the associated variable names. The (e) in parentheses following the specification of the residual variances of the five indicators, v01–v05, constrains them to be equal, yielding the essentially parallel model [EssentiallyParallel.inp].

```
! Essentially Parallel model
MODEL: f by v01-v05* (b);  ! Common factor loading
 f@1;                      ! Fix the factor variance to 1
 v01-v05 (e);              ! Common indicator residual variance
```

The strictly parallel model equates the indicator loadings (b), residual variances (e), and intercepts (a) [StrictlyParallel.inp].

```
! Strictly Parallel model
MODEL: f by v01-v05* (b);  ! Common factor loading
 f@1;                      ! Fix the factor variance at 1
 v01-v05 (e);              ! Common residual variance
 [v01-v05] (a);            ! Common intercept
```

Tau-equivalent models permit item-specific residual variances while Parallel models require a common residual variance for all items. "Essentially" Parallel or Tau-equivalent models permit item-specific intercepts, while "Strictly" Parallel or Tau-equivalent models require equal intercepts.

2.3 A Strictly Congeneric Model

Another potential simplification of the congeneric model is to restrict the expected values or predicted means of the indicators to be proportional to the factor loadings, given an estimated overall latent mean. This strictly congeneric model can be obtained by freely estimating the latent mean, α, while fixing the indicator intercepts to zero, $\nu_j = 0$. The strictly congeneric model also has $2J + 1$ parameters and is nested within the traditional congeneric model. Like the strictly tau-equivalent and parallel models when compared to their essentially tau-equivalent and parallel counterparts, the strictly congeneric model explicitly incorporates the mean structure of the data into the measurement evaluation process. The strictly congeneric model can be obtained by freely estimating the latent variable mean (which is fixed to zero by default in Mplus) and fixing the indicator intercepts to zero [StrictlyCongeneric.inp].

```
! Strictly Congeneric model
MODEL: f by v01-v05*;      ! Indicator-specific factor loadings
 f@1;                      ! Fix the factor variance at 1
 [f];                      ! Estimate the latent mean
 [v01-v05@0];              ! Fix the indicator intercepts at 0
```

2.4 Item Reliability

The reliability of individual indicators can be assessed by examining the proportion of an indicator's variance that can be explained by the common factor. The reliabilities of the individual indicators can be expressed as $R_j^2 = \lambda_j^2 / \left(\lambda_j^2 + \theta_{jj}\right)$. These are available in Mplus by requesting the STDYX fully standardized R-square estimates. If the parallel model fits the data, the indicators can further be considered to be equally reliable, $R^2 = \lambda^2/(\lambda^2 + \theta)$, because the loadings are all equal, as are the residual variances.

With the usual congeneric model, a weighted average of the item reliabilities is sometimes known as the average variance extracted (AVE).

$$AVE = \frac{\sum_{j=1}^{J} \lambda_j^2}{\sum_{j=1}^{J} \lambda_j^2 + \sum_{j=1}^{J} \theta_j}$$

AVE is not a traditional reliability coefficient, but instead expresses the average proportion of the item variance that can be accounted for by a single underlying latent factor. Values of AVE near or above 0.50 are often seen as reflecting a well-measured construct. The following Mplus syntax can be used to estimate the item reliabilities and AVE as new parameters [AverageVarianceExtracted.inp]. Although the item reliabilities are available in Mplus' default standardized results, it is necessary to explicitly re-compute them to obtain their associated bootstrap confidence intervals.

```
TITLE: Item reliabilities and Average Variance Extracted (AVE)
DATA: FILE IS selectedExampleData.dat;
VARIABLE: NAMES ARE v01 v02 v03 v04 v05 v11 v12 v13 v14 v15;
USEVARIABLES ARE v01-v05; ! Select variables for analysis
ANALYSIS: BOOTSTRAP=1000; ! Bootstrap
MODEL: f@1;               ! Fix the factor variance to 1
 f by v01-v05* (b1-b5);   ! Indicator factor loadings
 v01-v05 (e1-e5);         ! Label the residual variances
MODEL CONSTRAINT:         ! Model constraint
 NEW(IR1-IR5 AVE);        ! New parameters
 DO (1,5) IR#=b#^2/(b#^2+e#);    ! Item reliabilities
 AVE=(b1^2+b2^2+b3^2+b4^2+b5^2)/ ! AVE
    ((b1^2+b2^2+b3^2+b4^2+b5^2)+ !
      e1+e2+e3+e4+e5);    !
OUTPUT: STDYX;            ! Standardized results
 CINTERVAL(BCBOOTSTRAP);  ! Bootstrap confidence intervals
```

2.5 The Congeneric Model with Equally Reliable Indicators

Although the congeneric model estimates indicator-specific factor loadings, it is nevertheless also potentially advantageous for the indicators to be equally reliable. This would be seen if the relationship between the true score variance and the error variance were the same for each indicator, $R_1^2 = R_2^2 \cdots = R^2$. This equally-reliable congeneric model has the same number of parameters as the essentially tau-equivalent model, $2J + 1$, and therefore can be more parsimonious than the traditional congeneric model while potentially still fitting the data reasonably well. With latent variable modeling software that permits nonlinear constraints such as Mplus, it can be convenient to constrain the reliabilities of the indicators to be equal, $R^2 = \lambda_j^2 / \left(\lambda_j^2 + \theta_{jj}\right)$. The congeneric model with equally reliable indicators can be estimated using Mplus' MODEL CONSTRAINT: command to create a new parameter R which is the common indicator reliability [EquallyReliableCongeneric.inp].

```
! Congeneric model with equally reliable indicators
MODEL: f by v01-v05* (b1-b5); ! Factor loadings
 f@1;                        ! Fix the factor variance to 1
 v01-v05 (e1-e5);            ! Label the indicator residual variances
MODEL CONSTRAINT: NEW(R);    ! Equally reliable indicators
 DO (1,5) R=b#^2/(b#^2+e#);  !
OUTPUT: STDYX CINTERVAL;     ! Standardized results & confidence
                               intervals
```

The DO statement in the MODEL CONSTRAINT: command above is equivalent to having five separate model constraint equations.

```
R=b1^2/(b1^2+e1);
R=b2^2/(b2^2+e2);
R=b3^2/(b3^2+e3);
R=b4^2/(b4^2+e4);
R=b5^2/(b5^2+e5);
```

The estimated common item reliability of 0.481 with a 95% confidence interval of [0.421, 0.540] is given in the Mplus output.

```
New/Additional Parameters
                    Estimate       S.E.  Est./S.E.    P-Value
    R                  0.481      0.030     15.839      0.000
```

Model fit statistics for many of the models for the example data are given in Table 1.4. The congeneric models fit this data much better than the tau-equivalent and parallel models. The strictly tau-equivalent model and the two parallel models give completely nonsensical results for the CFI, TLI, and SRMR. Indeed, among the models in Table 1.1, the standard independence model with freely estimated variances and intercepts and all covariances set to zero ($\chi^2(10) = 402.737$) only provide an appropriate incremental fit baseline for the essentially tau-equivalent model, the essentially congeneric model, and the congeneric model with one residual covariance (see Table 1.3). The CFI and TLI values from the default Mplus output for the remaining models are not properly scaled and have an uncertain or improper interpretation (Widaman & Thompson, 2003).

As previously outlined, the parameters estimated from many of the models are equated in various ways to obtain more or less parsimonious empirical accounts for the given data (see Table 1.5). The most restricted model, the strictly parallel model, estimates a common factor loading, a

Table 1.4 Fit of different models to the example data.

	Par	LL	Chi-SQ	DF	RMSEA	CFI	TLI	SRMR
Congeneric with 1 residual covariace	16	−3734.474	4.184	4	0.014	**1.000**	**0.999**	0.015
Essentially congeneric	15	−3739.012	13.260	5	0.081	**0.979**	**0.958**	0.027
Equally reliable congeneric	11	−3740.134	15.504	9	0.054	0.983	0.982	0.034
Strictly congeneric	11	−3766.389	68.014	9	0.162	0.850	0.833	0.191
Essentially tau-equivalent	11	−3804.328	143.893	9	0.245	**0.657**	**0.618**	0.265
Strictly tau-equivalent	7	−4400.378	1335.993	13	0.638	0	−1.591	1.817
Essentially parallel	7	−4072.212	679.660	13	0.453	0	−0.306	1.667
Strictly parallel	3	−4584.152	1703.539	17	0.630	0	−1.526	2.894

Par, number of parameters; LL, log likelihood; Chi-SQ, Chi-square; DF, degrees of freedom; RMSEA, root mean square error of approximation; CFI, comparative fit index; TLI, Tucker–Lewis index; SRMR, standardized root mean square residual.
Only the **bolded CFI** and **TLI** values are valid.

Table 1.5 Model parameters from selected models for the example dataset.

	Essentially congeneric	Strictly congeneric	Essentially tau-equivalent	Strictly tau-equivalent	Essentially parallel	Strictly parallel
Factor mean	0	7.65	0	0	0	0
Factor loadings						
V01	1.982	2.341	*2.620*	*2.424*	*4.287*	*2.779*
V02	6.914	4.177	*2.620*	*2.424*	*4.287*	*2.779*
V03	2.341	2.645	*2.620*	*2.424*	*4.287*	*2.779*
V04	4.369	4.503	*2.620*	*2.424*	*4.287*	*2.779*
V05	6.423	3.926	*2.620*	*2.424*	*4.287*	*2.779*
Intercepts						
V01	17.939	0	17.938	*20.097*	17.938	*26.829*
V02	31.695	0	31.695	*20.097*	31.695	*26.829*
V03	20.257	0	20.257	*20.097*	20.257	*26.829*
V04	34.475	0	34.475	*20.097*	34.475	*26.829*
V05	29.781	0	29.781	*20.097*	29.781	*26.829*
Residual variances						
V01	4.536	3.998	2.981	10.680	29.861	*83.135*
V02	42.527	56.500	68.687	210.186	29.861	*83.135*
V03	6.936	6.381	6.530	5.410	29.861	*83.135*
V04	16.641	17.596	23.941	235.411	29.861	*83.135*
V05	53.002	63.934	73.396	171.701	29.861	*83.135*
No. of parameters	15	11	11	7	7	3

Parameters in *italics* are equated across items.

common intercept, and a common residual variance for all of the items. The essentially parallel model equates the loadings and the residual variances but estimates item-specific intercepts. The strictly tau-equivalent model equates the loadings and intercepts but allows separate residual variances for each item. The essentially tau-equivalent model has equal factor loadings but permits item-specific intercepts and residual variances. The essentially congeneric model has item-specific loadings, intercepts, and residual variances. In most of these models, the factor mean is fixed to 0 in order to identify the mean structure of the data. However, the strictly congeneric model estimates a factor mean and fixes the item intercepts to zero, while freely estimating the factor loadings and residual variances.

3 Composite Reliability

A common conceptual description of reliability is how well the sum of the item scores reflects the desired construct measured by the items. Most commonly, such a unit-weighted composite is used to measure the attribute of interest. Sometimes, the mean is used rather than the sum, or the items may be combined using theoretically or empirically defined differential weights.

3.1 Coefficient Alpha

One of the most common uses of true score theory is to assess the reliability of a composite score created by combining various subtests, part scores, indicators, etc., to create a single summary score for the measure. In an effort to assess the so-called internal consistency reliability of a composite mean or sum score, a number of different reliability coefficients have been employed, but by far the most common has been Cronbach's Coefficient Alpha. Alpha is included in many commercial general-purpose statistical software packages such as SAS, SPSS, and Stata, and is thus almost ubiquitously available to, and utilized by, researchers despite continuing concerns about its appropriateness and suggestions of alternative reliability measures. Although it is generally much simpler to use the built-in procedures in standard software packages, Alpha is a simple function of the variances and covariances of the indicators, and can be readily calculated from the indicator variances and covariances using a statistical modeling program such as Mplus.

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J} s_{jj}}{\sum_{j=1}^{J}\sum_{j'=1}^{J} s_{jj'}}\right)$$

The last term of the formula within the parentheses is the ratio of the sum of the indicator variances to the sum of the indicator variances and covariances. Equivalently, it is equal to the trace of the item covariance matrix divided by the sum of all of the elements of the matrix. For example, using data for five items from Peters (2014), it is possible to calculate Alpha based on the $J = 5$ variances, labeled v1–v5, and the $J(J-1)/2 = 5(5-1)/2 = 10$ unique covariances, labeled c1–c10, on the MODEL: command of the Mplus syntax given below. The MODEL CONSTRAINT: command is then used to create a set of new parameters consisting of the sum of the variances, v, and the sum of the unique covariances, c, which are then used to calculate the Alpha coefficient. Because the covariance matrix is symmetric, the sum of the unique covariances must be multiplied by 2 to account for those both above and below the diagonal of the covariance matrix [CoefficientAlpha.inp]. The CINTERVAL option on the OUTPUT command provides (delta-method) confidence intervals for the newly calculated parameters such as alpha as well as for the other estimated model parameters.

```
TITLE: Coefficient Alpha
DATA: FILE IS selectedExampleData.dat;
VARIABLE: NAMES ARE v01 v02 v03 v04 v05 v11 v12 v13 v14 v15;
 USEVARIABLES ARE v01-v05;          ! Select variables for analysis
MODEL: v01-v05 (v1-v5);             ! J Variances
v01-v05 WITH v01-v05 (c1-c10);      ! J(J-1)/2 Covariances
MODEL CONSTRAINT: NEW(j v c alpha); ! Model constraints
 j=5;                               ! Number of items
 v=sum(v1-v5);                      ! Sum of the variances
 c=sum(c1-c10);                     ! Sum of the covariances
 alpha=(j/(j-1))*(1-(v)/(v+2*c));   ! Calculate Coefficient Alpha
OUTPUT: STDYX CINTERVAL;            ! Standardized results & CI'S
```

Again, an important advantage of using Mplus to calculate Alpha is the handling of missing data through FIML techniques, allowing the use of all available data in cases where some of the data may be missing. The usual procedures for calculating Alpha in general-purpose statistical software packages generally use listwise deletion and can therefore give potentially biased estimates. Listwise deletion discards all data from individuals who have any missing items. Another advantage is that Mplus directly provides a (delta-method) standard error and can give related confidence intervals for the estimated value of Alpha. Other methods are available for obtaining standard errors and confidence intervals for Alpha and other reliability coefficients (see Kelley & Pornprasertmanit, 2016), but most standard statistical package estimation procedures give only a point estimate for Alpha. The Mplus output provides the estimated Alpha coefficient.

```
                    Estimate       S.E.  Est./S.E.    P-Value
...
New/Additional Parameters
    ALPHA              0.755      0.020     38.233      0.000
```

The estimated Alpha coefficient for Peters' example data was 0.755 with a standard error of 0.020 and a 95% confidence interval of [0.716, 0.793]. This confidence interval is an approximation using the so-called delta method based on asymptotic statistical theory and can be slightly biased and even give out-of-range values when reliability estimates are close to the boundaries of zero or one.

3.2 Bootstrapping Coefficient Alpha

Delta method confidence intervals are convenient and can be useful in many cases, but bootstrap confidence intervals will generally be more accurate and precise. Mplus allows bootstrapping of the model and derived parameters, including Coefficient Alpha. This can be accomplished by using the following Mplus syntax. The bootstrap option on the ANALYSIS: command requests 1000 bootstrap samples and the OUTPUT: command requests bias-corrected bootstrap confidence intervals [BootstrappingAlpha.inp]. Traditional percentile bootstrap confidence intervals (without the bias correction procedure) can be obtained by instead using CINTERVAL(BOOTSTRAP) on the OUTPUT: command.

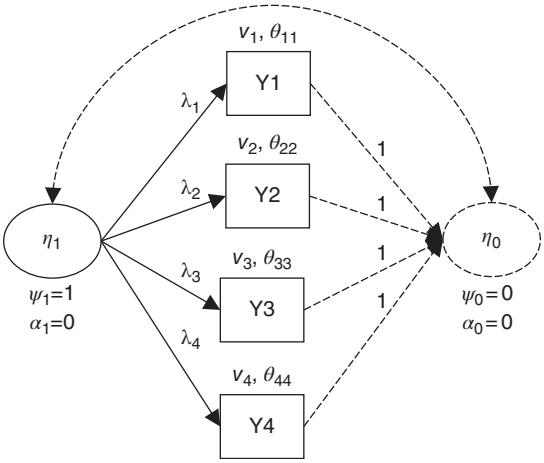
```
ANALYSIS: BOOTSTRAP=1000;       ! 1000 bootstrap samples
...
OUTPUT: CINTERVAL(BCBOOTSTRAP); ! Bootstrap confidence intervals
```

The estimate of Alpha (0.755) has a bootstrap standard error of 0.019 and a 95% bias-corrected bootstrap confidence interval of [0.713, 0.788].

3.3 The Reliability Index for Congeneric Data

Under certain conditions, Alpha is treated by some as a lower bound estimate of the population reliability coefficient. If these conditions are not satisfied, Alpha is an estimate of the population Coefficient Alpha, not an estimate of the population reliability coefficient. Instead of using Alpha, some have noted that the correlation between the latent variable η_1 and a unit-weighted linear composite score η_0 with an appropriately structured model (see Figure 1.4), also known as the reliability

Figure 1.4 Model to obtain the reliability index.



index, can be obtained from the latent variable correlations often provided as supplemental output by latent variable modeling software such as Mplus. The reliability index or estimated factor-composite correlation for a congeneric one-factor model is defined as the square root of the reliability, and the estimated reliability can be obtained by squaring the reliability index.

This can be accomplished using the following Mplus syntax [ReliabilityIndex.inp] which models the composite score c along with the common factor f.

```
! Reliability index for congeneric data
MODEL: f by v01-v05*;       ! Indicator-specific factor loadings
 f@1;                       ! Fix the factor variance to 1
 c BY;                      ! Define the composite variable
 c@0;                       ! Set the composite residual variance to 0
 c ON v01-v05@1;            ! Create an equally-weighted composite
OUTPUT: STDYX TECH4;        ! Standardized latent variable
                              correlations
```

The reliability index can be obtained from the TECH4 section of the Mplus output

```
TECHNICAL 4 OUTPUT
          ESTIMATED CORRELATION MATRIX FOR THE LATENT VARIABLES
               F          C          V01        V02             V03
               _______    _______    _______    _______         _______
 F             1.000
 C             0.893      1.000
```

Although this analysis gives a point estimate and standard error for the reliability index, it does not provide a confidence interval. For Peters' example data, the estimated correlation between the latent variable for the congeneric model and an equally-weighted composite, or the reliability index, was 0.893, yielding an estimated reliability of $0.893^2 = 0.797$.

3.4 Alternate Estimation of Coefficient Alpha

Models for the estimation of Coefficient Alpha based on the variances and covariances can become challenging to specify, especially as the number of indicators becomes large. Other methods for obtaining Alpha such as Hoyt's (1941) ANOVA approach are also available and produce identical results. Hoyt's method uses the estimated mean square values from a person by item repeated measures or mixed effects model.

$$\alpha = \frac{MS_{persons} - MS_{residual}}{MS_{persons}}$$

Another simple method is to estimate the essentially parallel true score model, along with a latent composite variable to get the reliability index for Alpha, as well as the Alpha estimate as a derived parameter using Mplus. The model has a single latent factor f with a common factor loading and residual variance, along with an equally-weighted latent composite variable s. In this situation, the formula for estimating Alpha is

$$\alpha = \frac{\lambda^2}{\lambda^2 + \theta/J}$$

where λ^2 is the squared common factor loading, θ is the common residual variance, and J is the number of indicators. Because the ANOVA and fitted parallel SEM models both provide estimates of α, we can equate the two expressions

$$\alpha = \frac{\lambda^2}{\lambda^2 + \theta/J} = \frac{\left[\lambda^2 + (\theta/J)\right] - (\theta/J)}{\lambda^2 + \theta/J} = \frac{MS_{persons} - MS_{residual}}{MS_{persons}}$$

The residual mean square from the ANOVA model is equal to the scaled residual variance $MS_{residual} = \theta/J$ from the fitted essentially parallel SEM model. The persons mean square from the ANOVA model is equal to the sum of the true score variance and the scaled residual variance $MS_{persons} = \lambda^2 + \theta/J$ from the fitted parallel SEM model. The estimated true score variance from the fitted parallel SEM model is equal to the difference between the mean square for persons and the residual mean square $\lambda^2 = MS_{persons} - MS_{residual}$ from the ANOVA model. The essentially parallel model can be fitted and the necessary calculations for Alpha can be accomplished with the following Mplus syntax [AlphaFromParallel.inp].

```
! Alpha from the essentially parallel model
MODEL: f@1;            ! Fix the factor variance to 1
 f by v01-v05* (b);    ! Common factor loading
 v01-v05 (e);          ! Common residual variance
 s BY; s@0;            ! Sum of indicators
 s ON v01-v05@1;       ! Equal weights
MODEL CONSTRAINT:      ! Model constraint
 NEW(J alpha);         ! New parameters
 J=5;                  ! Number of indicators
 alpha=b^2/(b^2+e/J); ! Coefficient alpha
OUTPUT: tech4;
```

The new parameter output gives the estimated Alpha coefficient.

```
                    Estimate       S.E.  Est./S.E.    P-Value
New/Additional Parameters
    ALPHA              0.755      0.025     30.776      0.000
```

The reliability index from the TECH4 output for this model gives the correlation between the parallel model latent factor f and the latent composite score s of 0.869. This correlation is the reliability index, giving the estimated Alpha coefficient when squared ($0.869^2 = 755$).

```
            ESTIMATED CORRELATION MATRIX FOR THE LATENT VARIABLES
               F             S

               _____         _____
 F               1.000
 S               0.869         1.000
```

Notice that the reported standard error for Alpha calculated from estimating the (ill-fitting) parallel model (0.025) is larger than for Alpha from fitting the full (saturated) covariance structure model (0.020), which differs in turn from the bootstrap standard error (0.019). In general, the bootstrap standard error and confidence interval are preferred.

3.5 Coefficient Omega

When the single-factor congeneric model fits the data well, a model-based reliability coefficient known as Omega is frequently recommended.

$$\omega = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \left(\sum_{j=1}^{J} \theta_{jj}\right)}$$

Omega can be estimated using the following Mplus syntax [CoefficientOmega.inp].

```
! Coefficient Omega
MODEL: f by v01-v05* (b1-b5);      ! Indicator-specific factor
                                     loadings
 f@1;                              ! Fix the factor variance to 1
 v01-v05 (e1-e5);                  ! Label the residual variances
MODEL CONSTRAINT: NEW(b e omega); ! Calculate Coefficient Omega
 b=sum(b1-b5);                     ! Sum of factor loadings
 e=sum(e1-e5);                     ! Sum of residual variances
 omega=b^2/(b^2+e);                ! Coefficient Omega
OUTPUT: STDYX CINTERVAL;           ! Confidence intervals
```

The Mplus output gives the estimated Omega coefficient.

```
New/Additional Parameters
                    Estimate       S.E.  Est./S.E.    P-Value
    OMEGA                0.797      0.022     36.657      0.000
```

With a well-fitting factor model having no correlated errors, Omega is an appropriate estimate of the population reliability coefficient. For Peters' example data, the estimated Omega coefficient was 0.797 with a (delta method) standard error of 0.022 and a 95% confidence interval [0.754, 0.840].

3.6 Bootstrapping Omega

Mplus permits bootstrapping of the additional model parameters such as the estimate of Omega. Again, this can be accomplished by using the bootstrap option on the Mplus ANALYSIS: command, and asking for bias-corrected bootstrap confidence intervals [BootstrappingOmega.inp].

```
ANALYSIS: BOOTSTRAP=1000;       ! 1000 bootstrap samples
...
OUTPUT: CINTERVAL(BCBOOTSTRAP); ! Bootstrap confidence intervals
```

The bootstrap analysis gave the point estimate (0.797) and bootstrap standard error (0.022) for Omega, with a 95% bias-corrected bootstrap confidence interval of [0.747, 0.834].

3.7 Jointly Estimating and Comparing Alpha and Omega

We have estimated Coefficient Alpha by itself using the conventional variance and covariance formula and also using the essentially parallel model. If the single-factor congeneric model fits the data well, it can also be useful to estimate both Alpha and Omega in the same computer analysis run, and be able to test the difference between the two coefficients. The numerator of the conventional formula for Omega (repeated here) is the estimated true score variance, and the denominator is the estimated total variance combining the estimated true score variance and the estimated error variance.

$$\omega = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \left(\sum_{j=1}^{J} \theta_{jj}\right)}$$

The numerator of the rightmost quantity in the usual formula for Coefficient Alpha is the sum of the observed indicator variances and the denominator of this is the estimated total variance (see here).

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J} S_{jj}}{\sum_{j=1}^{J}\sum_{j'=1}^{J} S_{jj'}}\right)$$

For a well-fitting single-factor congeneric model the total variance can be approximated by

$$\left(\sum_{j=1}^{J}\lambda_j\right)^2 + \left(\sum_{j=1}^{J}\theta_{jj}\right) \approx \sum_{j=1}^{J}\sum_{j'=1}^{J} S_{jj'}$$

In addition, in such a model the variance of a given indicator is a function of the associated true score and error components.

$$S_{jj} \approx \lambda_j^2 + \theta_{jj}$$

Combining these results,

$$\alpha \approx \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J}\left(\lambda_j^2 + \theta_{jj}\right)}{\left(\sum_{j=1}^{J}\lambda_j\right)^2 + \left(\sum_{j=1}^{J}\theta_{jj}\right)}\right)$$

This allows estimation of Omega along with the approximation of Alpha using the following Mplus syntax.

```
TITLE: Jointly estimate and compare alpha and omega
DATA: FILE IS selectedExampleData.dat;
VARIABLE: NAMES ARE v01 v02 v03 v04 v05 v11 v12 v13 v14 v15;
 USEVARIABLES ARE v01-v05; ! Select variables for analysis
ANALYSIS: BOOTSTRAP=1000;  ! Bootstrap
MODEL: f@1;                ! Fix the factor variance to 1
 f by v01-v05* (b1-b5);    ! Common factor loading
 v01-v05 (e1-e5);          ! Common residual variance
MODEL CONSTRAINT: NEW(J b  ! Model constraint
 e v1-v5 v t omega alpha); ! New parameters
 J=5;                      ! Number of indicators
 b=sum(b1-b5);             ! Sum of loadings
 e=sum(e1-e5);             ! Sum of residual variances
 t=b^2+e;                  ! Estimated total variance
 DO (1,5) v#=b#^2+e#;      ! Model implied indicator variances
 v=sum(v1-v5);             ! Sum of indicator variances
 omega=b^2/t;              ! Omega coefficient
 alpha=(j/(j-1))*(1-v/t);  ! Alpha Coefficient
 NEW(D); D=omega-alpha;    ! Difference between alpha and omega
OUTPUT: CINTERVAL(BCBOOTSTRAP); ! Confidence intervals
```

This uses bootstrapping to estimate Alpha and Omega, and their bootstrap standard errors, and calculates the difference between Alpha and Omega and gives an associated significance test.

```
                    Estimate       S.E.  Est./S.E.    P-Value
New/Additional Parameters
    OMEGA              0.797      0.022     36.646      0.000
    ALPHA              0.755      0.019     38.835      0.000
    D                  0.042      0.006      6.675      0.000
```

The 95% bias-corrected bootstrap confidence intervals were [0.747, 0.834] for Omega and [0.714, 0.788] for Alpha. In this case, we find that Omega significantly exceeds the approximated Alpha coefficient.

3.8 Omega If an Item Is Deleted

A common task in measure development and revision is to locate any poorer or less optimal items that might be removed in order to improve the measure or to shorten it without substantially reducing its quality. As with reliability generally, the most common approach to this problem is to examine the value of Cronbach's Alpha if a given item were deleted. Because Alpha is dependent on often unmet assumptions, and sometimes may not be a good estimate of a measure's reliability, Raykov (2007, 2008) proposes using Omega if an item were deleted instead. This can be accomplished with the following Mplus syntax [OmegaIfDeleted.inp].

```
! Omega if item deleted
MODEL: f by v01-v05* (b1-b5); ! Indicator-specific factor loadings
 f@1;                         ! Fix the factor variance to 1
 v01-v05 (e1-e5);             ! Label the residual variances
MODEL CONSTRAINT:             ! Model constraints
NEW(b e omega d1-d5);         ! New parameters
 b=sum(b1-b5);                ! Sum of loadings
 e=sum(e1-e5);                ! Sum of residual variances
 omega=b^2/(b^2+e);           ! Calculate Omega
 DO (1,5) d#=(b-b#)^2/        ! Omega if item # deleted
  ((b-b#)^2+(e-e#));          !
OUTPUT: CINTERVAL;            ! Confidence intervals
```

The estimated Omega coefficients if a given item were deleted are given as new parameters in the associated Mplus output.

```
New/Additional Parameters
                    Estimate       S.E.  Est./S.E.    P-Value
    OMEGA              0.797      0.022     36.657      0.000
    D1                 0.771      0.025     31.115      0.000
    D2                 0.738      0.029     25.231      0.000
    D3                 0.769      0.025     30.455      0.000
    D4                 0.745      0.028     26.688      0.000
    D5                 0.775      0.026     29.331      0.000
```

In the example data, the overall Omega coefficient was 0.797 while the estimated Omega reliability coefficients when separately omitting items 1 through 5 were 0.771, 0.738, 0.769, 0.745, and

0.775, respectively. Standard errors are provided and confidence intervals can be requested for each of these. In addition, bootstrapping can be utilized to obtain bootstrap confidence intervals.

3.9 Coefficient Rho

The fit of the congeneric model for Peters' data was $\chi^2(5) = 13.260$, $p = 0.0211$, CFI = 0.979, TLI = 0.958, RMSEA = 0.081, and SRMR = 0.027. Though the RMSEA is a little high, this is generally considered quite good model fit, and a reasonable basis for using Omega as a reliability estimate. However, allowing a residual covariance between items 2 and 4 improves model fit significantly ($\chi^2(4) = 4.184$, $p = 0.3817$, CFI = 1.000, TLI = 0.999, RMSEA = 0.014, SRMR = 0.015). Where there are nontrivial correlated errors (see Figure 1.5), the composite reliability can be misestimated with reliability coefficients, such as Alpha and Omega, that fail to take the error correlations into account. The appropriate reliability coefficient in this case is known as rho.

$$\rho = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \left(\sum_{j=1}^{J} \theta_{jj}\right) + 2\left(\sum_{j=1}^{J} \theta_{jj'}\right)}$$

The point estimate of rho for this model was 0.756 with a standard error of 0.030 and a delta-method 95% confidence interval of [0.698, 0.815]. With 1000 bootstrap samples, the 95% bias-corrected bootstrap confidence interval was [0.687, 0.807]. The estimated reliability index for the model with the error correlation was 0.870 yielding a corresponding reliability estimate of 0.757. With a single nonzero residual covariance, this model can be estimated with the following Mplus syntax [CoefficientRho.inp].

```
! Coefficient Rho
MODEL: f by v01-v05* (b1-b5);    ! Indicator-specific factor loadings
 f@1;                            ! Fix the factor variance to 1
 v01-v05 (e1-e5);                ! Residual variances
 v02 WITH v04 (c);               ! Residual covariance (correlated
                                   error)
MODEL CONSTRAINT: NEW(b e rho); ! Calculate Raykov's rho
 b=sum(b1-b5);                   ! Sum of loadings
 e=sum(e1-e5);                   ! Sum of residual variances
 rho=b^2/(b^2+e+2*c);            ! Calculate rho
OUTPUT: CINTERVAL;               ! Confidence intervals
```

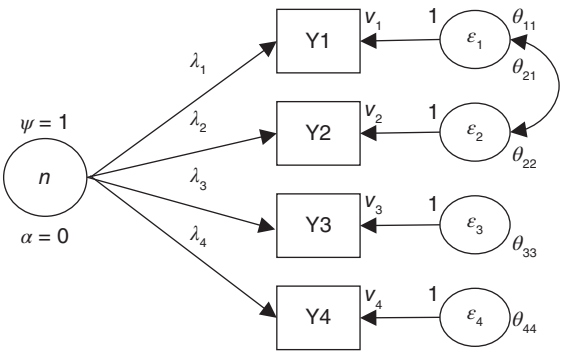


Figure 1.5 Single-factor model with a residual covariance or correlated error θ_{21}.

The Mplus output provides the estimated Rho coefficient.

```
New/Additional Parameters
                    Estimate       S.E.  Est./S.E.    P-Value
    RHO                0.756      0.030     25.519      0.000
```

3.10 Maximal Reliability and Optimally Weighted Indicators

To this point, our discussion of composite reliability has concerned an unweighted, unit-weighted, or equally weighted linear combination such as the sum or mean of the indicators. Methodologists and researchers have sometimes wondered whether it would be possible to find a differentially weighted linear combination of the indicators that could be more reliable than an equally weighted composite scale. When a congeneric model fits the data well, the optimal weights for such a linear combination are $w_j = \lambda_j/\theta_j$ and the maximal reliability of this optimally weighted linear composite is:

$$\rho_{max} = \frac{\sum_{j=1}^{J} \lambda_j^2/\theta_j}{1 + \sum_{j=1}^{J} \lambda_j^2/\theta_j} = \frac{\sum_{j=1}^{J} w_j\lambda_j}{1 + \sum_{j=1}^{J} w_j\lambda_j}$$

As shown in Figure 1.6, the optimally weighted items are used to construct the phantom latent composite score. This analysis can be accomplished with the following Mplus syntax [MaximalReliability.inp]:

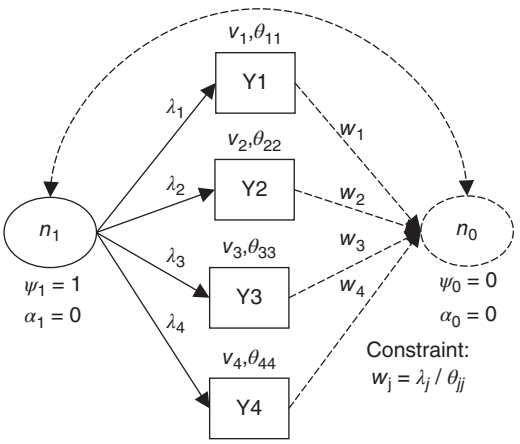


Figure 1.6 Model to obtain the reliability index for maximal reliability.

```
TITLE: Maximal Reliability and Maximal Reliability index
 With optimal indicator weights and bootstrapping
DATA: FILE IS selectedExampleData.dat;
VARIABLE: NAMES ARE v01 v02 v03 v04 v05 v11 v12 v13 v14 v15;
 USEVARIABLES ARE v01-v05; ! Select variables for analysis
ANALYSIS: BOOTSTRAP=1000; ! Bootstrap
MODEL:                    ! Model
```

```
 f by v01-v05* (b1-b5);   ! Indicator-specific factor loadings
 f@1;                     ! Fix the factor variance to 1
 v01-v05 (e1-e5);         ! Label the residual variances
 s BY; s@0;               ! Composite
 s ON v01-v05 (w1-w5);    ! Optimal weights
MODEL CONSTRAINT:         ! Model constraint
 NEW(r1-r5 r MR);         ! New variables
 DO (1,5) w#=b#/e#;       ! Optimal indicator weights
 DO (1,5) r#=b#*w#;       ! Ratios of squared loadings to error
 r=sum(r1-r5);            ! Sum of ratios
 MR=r/(1+r);              ! Maximal reliability
OUTPUT: TECH4;            ! Maximal reliability index
 CINTERVAL(BCBOOTSTRAP);  ! Bootstrap confidence intervals
```

The Mplus output provides the estimated maximal reliability coefficient, along with the optimal weights.

```
                   Estimate       S.E.  Est./S.E.    P-Value
S        ON
    V01                0.437      0.070      6.270      0.000
    V02                0.163      0.027      6.080      0.000
    V03                0.338      0.054      6.203      0.000
    V04                0.263      0.043      6.066      0.000
    V05                0.121      0.019      6.301      0.000
New/Additional Parameters
    MR                 0.825      0.018     46.682      0.000
ESTIMATED CORRELATION MATRIX FOR THE LATENT VARIABLES
              F             S
 F              1.000
 S              0.908         1.000
```

With Peters' data, the estimated maximal reliability was 0.825 with a standard error of 0.018 and a 95% confidence interval of [0.790, 0.859]. The estimated optimal weights for items 1 through 5 were 0.437, 0.163, 0.338, 0.263, and 0.121. The item weights and the maximal reliability estimate can also be bootstrapped, as described earlier. By using a phantom latent composite variable, it is also possible to obtain an estimate of the maximal reliability index. The estimated maximal reliability index, or the correlation of the latent variable with the optimally weighted composite, was 0.908, yielding a reliability estimate of 0.824 when squared. The discrepancy in the third decimal place between this estimate and the standard calculated value is due to a rounding error.

4 Standardized Reliability

When indicators are on different scales, or have arbitrary metrics, some recommend computing reliabilities after having standardized the indicators.

4.1 Standardized Alpha

When the indicator variances are highly disparate, some recommend computing Alpha using the correlation matrix rather than the covariance matrix. This is essentially equivalent to standardizing all of the indicators (transforming them to z-scores) and then calculating Alpha for the transformed scores (Falk & Savalei, 2011). Standardized Alpha is a direct function of the average inter-item correlation $\bar{r}$.

$$\alpha_{stand} = \frac{J\bar{r}}{1 + (J-1)\bar{r}}$$

Where there are J items, the average correlation is the mean of the $J(J-1)/2$ possible correlations between pairs of items. Standardized Alpha can be obtained by directly analyzing the correlation matrix with weighted least squares (WLS) using the following Mplus syntax [StandardizedAlpha.inp].

```
! Standardized Alpha WLS
ANALYSIS: MATRIX=CORRELATION;      ! Analyze correlations using WLS
MODEL:                             ! Model
 v01-v05 WITH v01-v05 (r1-r10);    ! Label the J(J-1)/2 correlations
MODEL CONSTRAINT: NEW(j r salpha); ! New Parameters
 j=5;                              ! Number of items
 r=mean(r1-r10);                   ! Average correlation
 salpha=j*r/(1+(j-1)*r);           ! Standardized alpha
OUTPUT: CINTERVAL;                 ! Confidence intervals
```

The Mplus output provides the average inter-item correlation and the estimated standardized Alpha coefficient.

```
New/Additional Parameters
    R                  0.481      0.030      16.233      0.000
    SALPHA             0.822      0.017      47.444      0.000
```

For Peters' data, the average inter-item correlation was 0.481 with a standard error of 0.030 and a 95% confidence interval of [0.423, 0.539]. The corresponding WLS standardized Alpha estimate was 0.822 with a standard error of 0.017 and a 95% confidence interval of [0.788, 0.856].

Maximum likelihood estimation of standardized Alpha can be obtained with the following Mplus Syntax. Here a counterpart latent variable is created for each observed indicator while fixing the indicator residual variances to zero and the latent variances to 1.0. The freely estimated factor loadings are the indicator standard deviations [StandardizedAlphaML.inp].

```
! Standardized Alpha ML
MODEL: f1 BY v01*;          ! Latent variable for each indicator
 f2 by v02*; f3 by v03*;    !
 f4 by v04*; f5 by v05*;    !
 v01-v05@0;                 ! Fix item residual variances to 0
 f1-f5@1;                   ! Standardize the latent indicators
```

```
 f1-f5 WITH f1-f5 (r1-r10); ! Correlations
MODEL CONSTRAINT:           ! Model constraints
 NEW(j r salpha);           ! New Parameters
 j=5;                       ! Number of items
 r=mean(r1-r10);            ! Average correlation
 salpha=j*r/(1+(j-1)*r);    ! Calculate standardized alpha
OUTPUT: CINTERVAL;          ! Confidence intervals
```

As with WLS estimation, the maximum likelihood estimate of standardized alpha was 0.822 with a standard error of 0.018 and a 95% confidence interval of [0.787, 0.857].

4.2 Standardized Omega

Standardized Omega is essentially the computation of Coefficient Omega with standardized indicators. As with Coefficient alpha, some recommend using standardized Omega when the original observed indicators have highly disparate variances. With maximum likelihood estimation, this is done by creating a first-order factor for each observed indicator where the single loading for each indicator is freely estimated as the indicator's standard deviation. The residual variances for the observed indicators are fixed to zero and loadings of the first-order factors on the common second-order factor are estimated.

Based on the parameters in Figure 1.7, the formula for estimating standardized Omega can be given as follows:

$$\omega_{stand} = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \left(\sum_{j=1}^{J} \psi_{jj}\right)}$$

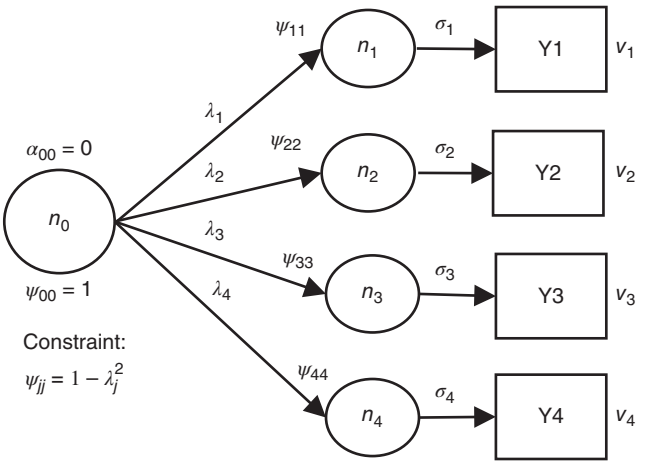


Figure 1.7 Standardized loadings and residual variances as model parameters.

where λ_j is the standardized factor loading and $\psi_{jj} = \left(1 - \lambda_j^2\right)$ is the standardized residual variance for item *J*. For Peters' data, this required providing a starting value (0.7) for the loadings. The residual variances for the first-order factors were constrained to fix the first-order factor variances to one. Standardized Omega was then calculated as usual using the following Mplus syntax. As with the maximum likelihood estimation of standardized Alpha, this requires creating unit-variance latent counterpart variables for each of the indicators and fixing the indicator residual variances to zero. Again, the freely estimated factor loadings approximate the indicator standard deviations [StandardizedOmega.inp].

```
! Standardized Omega
MODEL: f1 BY v01*;          ! Latent variable for each indicator
 f2 by v02*; f3 by v03*;    !
 f4 by v04*; f5 by v05*;    !
 v01-v05@0;                 ! Fix item residual variances to 0
 f@1;                       ! Fix the common factor variance to 1
 f by f1-f5*.7 (b1-b5);     ! Factor loadings (with starting values)
 f1-f5 (e1-e5);             ! Residual variances
MODEL CONSTRAINT:           ! Model constraints
 NEW(b e somega);           ! New parameters
 DO (1,5) e#=1-b#^2;        ! Standardize the latent indicators
 b=sum(b1-b5);              ! Sum of factor loadings
 e=sum(e1-e5);              ! Sum of residual variances
 somega=b^2/(b^2+e);        ! Standardized Omega
OUTPUT: STDYX CINTERVAL;    ! Confidence intervals
```

This produces model estimated factor loadings that are equal to the standardized loadings, and allows the standardized parameters to be labeled and used in the MODEL CONSTRAINT: command to create new parameters. In the Mplus output, we can verify that the unstandardized and standardized loadings are equal as shown below.

```
                    Estimate       S.E.  Est./S.E.    P-Value
 F        BY
    F1                 0.681      0.043     15.962      0.000
    F2                 0.727      0.040     18.397      0.000
    F3                 0.664      0.044     14.960      0.000
    F4                 0.731      0.039     18.617      0.000
    F5                 0.662      0.044     15.032      0.000
STDYX Standardization
 F        BY
    F1                 0.681      0.043     15.962      0.000
    F2                 0.727      0.040     18.397      0.000
    F3                 0.664      0.044     14.960      0.000
    F4                 0.731      0.039     18.617      0.000
    F5                 0.662      0.044     15.032      0.000
```

The standardized Omega estimate is also provided in the Mplus output.

	Estimate	S.E.	Est./S.E.	P-Value
New/Additional Parameters				
SOMEGA	**0.822**	0.018	46.443	0.000

For Peters' data, the ML estimate of standardized Omega was 0.822 with a standard error of 0.018 and a 95% confidence interval of [0.788, 0.857].

4.3 Standardized Rho

Like the model for standardized Omega, the model for standardized Rho requires estimating factor loadings and residual variances for standardized variables. In addition, a residual covariance ψ_{21} is also estimated to account for association among the indicators, which is not due to the influence of the common factor that is shared by all of the indicators. Where there are nontrivial correlated errors, a standardized composite reliability can be misestimated with reliability coefficients, such as standardized Alpha or standardized Omega, that fail to take the error correlations into account. The appropriate reliability coefficient, in this case, is known as standardized Rho.

Using the parameters from the model in Figure 1.8, the general formula for standardized Rho can be given as follows.

$$\rho_{stand} = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \left(\sum_{j=1}^{J} \psi_{jj}\right) + 2\left(\sum_{j=1}^{J} \psi_{jj'}\right)}$$

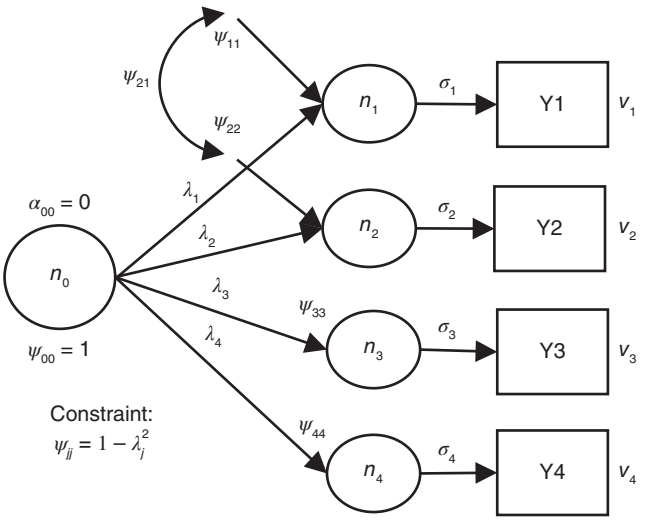


Figure 1.8 Standardized loadings and residual variances and covariances as model parameters.

The Mplus syntax for this analysis

```
! Standardized Rho
MODEL: f1 BY v01*;          ! Latent variable for each indicator
 f2 by v02*; f3 by v03*;    !
 f4 by v04*; f5 by v05*;    !
 v01-v05@0;                 ! Fix item residual variances to 0
 f@1;                       ! Fix the common factor variance to 1
 f by f1-f5*.7 (b1-b5);     ! Factor loadings (with starting values)
 f1-f5 (e1-e5);             ! Residual variances
 f2 WITH f4 (c);            ! Residual covariance
MODEL CONSTRAINT:           ! Model constraints
 NEW(b e srho);             ! New parameters
 DO (1,5) e#=1-b#^2;        ! Standardize the latent indicators
 b=sum(b1-b5);              ! Sum of factor loadings
 e=sum(e1-e5);              ! Sum of residual variances
 srho=b^2/(b^2+e+2*c);      ! Standardized Rho
```

The new parameter output gives the estimated standardized Rho coefficient.

```
                    Estimate       S.E.  Est./S.E.    P-Value
New/Additional Parameters
    SRHO               0.798      0.022     35.803      0.000
```

4.4 Standardized Maximal Reliability (Coefficient *H*)

Coefficient H is the estimated reliability of a linear combination of the item z-scores weighted by the ratio of the standardized factor loadings to their (standardized) uniquenesses (Hancock & Mueller, 2001; McNeish, 2018). The optimal standardized weights are equal to $w_j = \frac{\ell_j}{1-\ell_j^2}$ where ℓ_j are the standardized factor loadings. Maximal reliability or Coefficient H is then

$$H = 1/\left[1 + 1/\left(\sum_{j=1}^{J} \frac{\ell_j^2}{1-\ell_j^2}\right)\right] = 1/\left[1 + 1/\left(\sum_{j=1}^{J} w_j \ell_j\right)\right]$$

where maximal reliability is a function of the sum of the products of the standardized optimal weights and loadings.

Coefficient H can be obtained by using the following Mplus syntax [StandardizedMaximalReliability.inp]:

```
! Standardized maximal reliability
MODEL: f1 BY v01*;          ! Latent variable for each indicator
 f2 by v02*; f3 by v03*;    !
 f4 by v04*; f5 by v05*;    !
 v01-v05@0;                 ! Fix item residual variances to 0
 f@1;                       ! Fix the common factor variance to 1
```

```
 f by f1-f5*.7 (b1-b5);     ! Factor loadings (with starting values)
 f1-f5 (e1-e5);             ! Residual variances
MODEL CONSTRAINT:           ! Model constraints
 NEW(h1-h5 w1-w5 h CH);     ! New parameters
 DO (1,5) e#=1-b#^2;        ! Standardize the latent indicators
 DO (1,5) h#=b#^2/e#;       ! Variance ratios
 DO (1,5) w#=b#/e#;         ! Optimal indicator weights
 h=sum(h1-h5);              ! Sum of ratios
 CH=1/(1+1/h);              ! Standardized maximal reliability
OUTPUT: STDYX CINTERVAL;    ! Confidence intervals
```

This produces the following output with optimal weights w1–w5 and the standardized maximal reliability estimate of 0.825 with a standard error of 0.018 and a 95% confidence interval of [0.790, 0.859]. The weights are used to obtain an optimally weighted linear composite of the item z-scores having the estimated maximal reliability.

```
New/Additional Parameters
                    Estimate       S.E.  Est./S.E.    P-Value
    W1                 1.271      0.218      5.842      0.000
    W2                 1.545      0.273      5.664      0.000
    W3                 1.189      0.205      5.797      0.000
    W4                 1.570      0.278      5.651      0.000
    W5                 1.177      0.200      5.879      0.000
    H                  4.706      0.575      8.181      0.000
    CH                 0.825      0.018     46.681      0.000
```

"Standardized" Coefficient H is estimated for standardized items, and is identical to maximal reliability for unstandardized items. The formulas are algebraically equivalent, but both are used in the literature. However, the estimated optimal weights differ substantially. The standardized indicator weights are equal to the products of the unstandardized weights and the item standard deviations $w_j^{(S)} = w_j^{(U)}\sigma_j$.

5 Summary

For a unidimensional congeneric model, Table 1.6 summarizes reliability formulas for coefficient Omega and Coefficient H (maximal reliability) for both standardized and unstandardized scores. Table 1.6 also includes an approximation of Coefficient Alpha along with the AVE for both standardized and unstandardized scores, again derived from the model parameters of a single-factor congeneric model.

These coefficients and other statistics can be derived from the factor loadings and either the residual variances or the item standard deviations. Although we prefer to use SEM software that can derive new parameters from model estimation results, such as Mplus, general-purpose statistical software such as Stata, SPSS, or SAS that can carry out ordinary maximum likelihood factor analysis can be used to implement these formulas as well.

Table 1.6 Reliability and related formulas for original (unstandardized) and standardized scores.

	Standardized	Unstandardized
Congeneric model		
Omega	$\omega_{(s)} = \dfrac{\left(\sum_{j=1}^{J} \ell_j\right)^2}{\left(\sum_{j=1}^{J} \ell_j\right)^2 + \left(\sum_{j=1}^{J} 1-\ell_j^2\right)}$	$\omega = \dfrac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \left(\sum_{j=1}^{J} \theta_{jj}\right)}$
Omega if item deleted	$\omega_{j(s)} = \dfrac{\left[\left(\sum_{j=1}^{J} \ell_j\right) - \ell_j\right]^2}{\left[\left(\sum_{j=1}^{J} \ell_j\right) - \ell_j\right]^2 + \left[\left(\sum_{j=1}^{J} 1-\ell_j^2\right) - \left(1-\ell_j^2\right)\right]}$	$\omega_j = \dfrac{\left[\left(\sum_{j=1}^{J} \lambda_j\right) - \lambda_j\right]^2}{\left[\left(\sum_{j=1}^{J} \lambda_j\right) - \lambda_j\right]^2 + \left[\left(\sum_{j=1}^{J} \theta_{jj}\right) - \theta_{jj}\right]}$
Optimal weights	$w_{j(s)} = \dfrac{\ell_j}{1-\ell_j^2}$	$w_j = \dfrac{\lambda_j}{\theta_{jj}}$
Coefficient H	$H_{(s)} = 1/\left[1 + 1/\left(\sum_{j=1}^{J} \dfrac{\ell_j^2}{1-\ell_j^2}\right)\right]$	$H = \dfrac{\sum_{j=1}^{J} \lambda_j^2/\theta_{jj}}{1 + \sum_{j=1}^{J} \lambda_j^2/\theta_{jj}}$
Coefficient H via weights	$H_{(s)} = 1/\left[1 + 1/\left(\sum_{j=1}^{J} w_j \ell_j\right)\right]$	$H = \dfrac{\sum_{j=1}^{J} w_j \lambda_j}{1 + \sum_{j=1}^{J} w_j \lambda_j}$
Approximated alpha	$\widetilde{\alpha}_{(s)} \approx \dfrac{J}{J-1}\left(1 - \dfrac{J}{\left(\sum_{j=1}^{J} \ell_j\right)^2 + \left(\sum_{j=1}^{J} 1-\ell_j^2\right)}\right)$	$\widetilde{\alpha} \approx \dfrac{J}{J-1}\left(1 - \dfrac{\sum_{j=1}^{J}\left(\lambda_j^2 + \theta_{jj}\right)}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \left(\sum_{j=1}^{J} \theta_{jj}\right)}\right)$

Average variance extracted	$AVE_{(s)} = \frac{\sum_{j=1}^{J} \ell_j^2}{J}$	$AVE = \frac{\sum_{j=1}^{J} \lambda_j^2}{\sum_{j=1}^{J} \lambda_j^2 + \sum_{j=1}^{J} \theta_{jj}}$
Residual covariance		
Rho	$\rho_{(s)} = \frac{\left(\sum_{j=1}^{J} \ell_j\right)^2}{\left(\sum_{j=1}^{J} \ell_j\right)^2 + \left(\sum_{j=1}^{J} 1 - \ell_j^2\right) + 2\left(\sum_{j=1}^{J} \psi_{jj'}\right)}$	$\rho = \frac{\left(\sum_{j=1}^{J} \lambda_j\right)^2}{\left(\sum_{j=1}^{J} \lambda_j\right)^2 + \left(\sum_{j=1}^{J} \theta_{jj}\right) + 2\left(\sum_{j=1}^{J} \theta_{jj'}\right)}$

λ_j is the unstandardized factor loading for item j; θ_{jj} is the unstandardized residual variance for item j; $\theta_{jj'}$ is an unstandardized residual covariance between items j and j'; ℓ_j is the standardized factor loading for item j; $\psi_{jj'}$ is a standardized residual covariance between items j and j'.

Until recently, researchers have generally relied upon the wide availability of the Coefficient Alpha statistical procedures in general-purpose software packages to accomplish reliability assessment in their work, even though there is accumulating evidence that other approaches, such as Coefficient Omega, may be better in many cases. We note with some enthusiasm that this is now beginning to change with several new developments. Hayes and Coutts (2020) has provided a macro called OMEGA for SPSS and SAS that can estimate standardized and unstandardized Omega, along with the estimated Omega if an item were deleted. Hayes' OMEGA macro can also produce the reliability of the possible subsets and different combinations of items based on either Cronbach's Alpha or McDonald's Omega. SPSS (as of version 28) also now supports Coefficient Omega in their RELIABILITY procedure that has previously been limited to Coefficient Alpha. There is also a new Stata user-written procedure (Shaw, 2021) that can calculate the standardized Coefficient Omega. The R packages *userfriendlyscience, MBESS*, and *psych* also support Coefficient Omega and the freely available package *Jamovi* can estimate Coefficient Omega and Omega if an item is deleted.

The broader availability of these new resources represents an important advance toward improved practice in reliability assessment. Making these resources easier to use will also contribute to the wider diffusion of these improvements throughout the research community. Graphical user interfaces such as those in Hayes' OMEGA macro, SPSS, and Jamovi will undoubtedly encourage greater use of these new methods among practicing research scientists. Syntax options for procedures such as the OMEGA macro and SPSS can also simplify multiple and repeated analyses and make them more efficient. Hayes' (2020) macro and Shaw's user-written procedure are intended primarily for reliable uncomplicated general use by nonmethodologists.

In addition to wider dissemination of Omega-capable reliability software, there are a number of other important considerations. These include (i) building confidence intervals and developing inference methods for reliability coefficients, (ii) handling missing data, (iii) assessing reliability of differentially weighted item composites in addition to traditional unit-weighted ones, (iv) appropriately modeling certain forms of multidimensionality or correlated error, (v) applying extended reliability analysis to categorical (dichotomous and polytomous) data. The model-based approaches we have advocated and proposed can help address these concerns. Although we have not directly addressed reliability assessment for categorical data, considering it outside the scope of the current guide, model-based approaches are more readily extended to such cases than are more traditional methods and software.

With respect to confidence intervals for reliability coefficients, we note that users can request bootstrap standard errors and confidence intervals for Coefficient Omega with Hayes' OMEGA macro. However, most reliability software provides only point estimates of the reliability coefficients. Model-based methods using software like Mplus can straightforwardly provide bootstrap confidence intervals and can perform bootstrapped hypothesis testing.

With respect to missing data, most reliability software uses only listwise or pairwise deletion to address missing data. Modern missing data methods such as FIML or multiple imputations (MIs) are generally not supported. For example, the reliability procedures in SPSS, SAS, and Stata do not support the automated pooling of multiply-imputed analysis results. It is necessary instead to separately analyze the datasets and manually pool the results. Model-based approaches using almost any SEM software will support FIML estimation, and many also support MI with automated pooling.

Just as Cronbach's Alpha has easily been the most common approach to reliability assessment, the most common approach to scale construction and scoring has been the calculation of unit-weighted composite measures (sum scores) and the interpretation of reliability has been in terms

of how well the sum score can serve as a proxy for an underlying true latent variable. The centrality of traditional sum scores is as true of reliability assessment using Omega as it is when using Alpha. Maximal reliability uses model-based analysis to assess how well an optimally weighted linear combination of the item scores can serve as a proxy for the latent variable, yielding the values of the optimal weights as well as the associated reliability coefficient.

Whereas Coefficient Alpha assumed both unidimensionality and tau-equivalence, Coefficient Omega relaxes the tau-equivalence assumption but still requires unidimensionality. When it is possible to model the misfit of the unidimensional model in terms of one or more residual covariances or correlated errors, Coefficient Rho provides a consistent way to estimate reliability through treating the residual covariances as additional correlated error components. This requires augmenting the original congeneric model with the relevant residual covariance(s).

The estimated reliability coefficients from these last two main sections are given in Table 1.7. For these data, we note that bootstrapping provides asymmetric confidence intervals that are somewhat

Table 1.7 Estimated reliability coefficients for the example data.

Reliability coefficient	Estimate	Confidence interval 95%
Coefficient alpha		
(Unstandardized) alpha	0.755	[0.716, 0.793]
Bootstrapped alpha	0.755	[0.713, 0.788]
Coefficient omega		
(Unstandardized) omega	0.797	[0.754, 0.840]
Bootstrapped omega	0.797	[0.747, 0.834]
Omega (squared reliability index)	0.797	
Coefficient rho		
(Unstandardized) rho	0.756	[0.698, 0.815]
Bootstrapped rho	0.756	[0.687, 0.807]
Rho (squared reliability index)	0.757	
Maximal reliability with unstandardized indicators		
Maximal reliability	0.825	[0.790, 0.859]
Bootstrap maximal reliability	0.825	[0.783, 0.852]
Maximal reliability (squared reliability index)	0.824	
Standardized alpha		
Standardized alpha (WLS)	0.822	[0.788, 0.856]
Standardized alpha (ML)	0.822	[0.787, 0.857]
Bootstrap standardized alpha (ML)	0.822	[0.782, 0.853]
Standardized omega		
Standardized omega	0.822	[0.788, 0.857]
Bootstrap standardized omega	0.822	[0.781, 0.852]
Standardized omega (squared reliability index)	0.823	

(*Continued*)

Table 1.7 (Continued)

Reliability coefficient	Estimate	Confidence interval 95%
Standardized rho		
Standardized rho	0.798	[0.754, 0.842]
Bootstrap standardized rho	0.798	[0.749, 0.837]
Standardized rho (squared reliability index)	0.797	
Maximal reliability with standardized indicators		
Standardized maximal reliability	0.825	[0.790, 0.859]
Bootstrap standardized maximal reliability	0.825	[0.783, 0.852]
Standardized maximal reliability (squared reliability index)	0.824	

narrower than the default delta method intervals. Again, for these data, Omega was larger than Alpha and Maximal reliability was larger than Omega, and the squared reliability index values also equal those from the corresponding model-based reliability calculations. We note, however, that Mplus does not provide a direct way to obtain confidence intervals for, or bootstrap, the reliability index values.

All of these reliability coefficients assume the test components reflect a single dimension. The traditional Alpha coefficient is very widely used but is technically applicable only with essentially tau-equivalent indicators (having equal factor loadings or true score variance contributions) and no correlated errors. Omega, Rho, and Maximal Reliability are also applicable with congeneric indicators (having possibly varying factor loadings or true score variance contributions). Omega also assumes no correlated errors, but Rho accounts for error correlations by treating them as additional components of error. Alpha, Omega, and Rho represent the expected reliability of an equally weighted composite (sum or mean) of the indicators, whereas Maximal Reliability estimates the reliability of an optimally weighted linear combination of the indicators.

Standardized reliability coefficients provide estimates of the reliability of indicators which have been standardized or transformed into z-scores. Standardized Alpha is based on the correlation matrix and the average inter-item correlation, whereas Alpha is based on the covariance matrix and the average inter-item covariances. Standardized Omega is based on standardized factor loadings and the indicator uniqueness values, whereas Omega is based on unstandardized factor loadings and the indicator residual variances. In general, standardized Alpha and Omega differ from their unstandardized counterparts.

The Standardized Alpha and Omega coefficients are equal in this dataset, but we note that this is a coincidence; these coefficients need not be the same in general. We also note that the unstandardized and standardized Maximal Reliability coefficients are also the same. In this case, however, the two formulas are indeed equivalent and necessarily produce the same results in this and other data. We have also demonstrated how to obtain standardized Alpha, Omega, and Rho.

The ordinary maximum likelihood (delta method) standard errors and the resulting symmetric confidence intervals for reliability coefficients are based on generally unsatisfied asymptotic distribution assumptions, and it is usually preferable to use bootstrapping for statistical testing and forming more accurate asymmetric confidence intervals for the reliability coefficients. We have demonstrated bootstrapping for Alpha, Omega, and Rho, but not for Maximal Reliability or the

standardized reliability coefficients. Again, however, this is a direct extension of the other bootstrapping procedures.

In this guide, we have introduced and applied CTT, the true score model, and basic measurement reliability analysis. We showed how different true score models can be formulated as latent variable models, including various forms of the parallel, tau-equivalent, and congeneric models. We demonstrated the estimation of composite reliability coefficients including Alpha, Omega, Rho, and maximal reliability, along with the reliability index and versions of Alpha and Omega for standardized items. We addressed bootstrapping of these reliability coefficients for more accurate confidence intervals and how to appropriately calculate the expected (Omega) reliability if an item were deleted.

References

Falk, C. F., & Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of Personality Assessment*, *93*(5), 445–453. https://doi.org/10.1080/00223891.2011.594129

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, *66*(6), 930–944. https://doi.org/10.1177/0013164406288165

Hancock, G. R., & Muller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudek, K. G. Jöreskog, D. Sorbom, & S. du Toit (Eds.), *Structural equation modeling: Present and future: A Festschrift in honor of Karl Jöreskog*, (pp. 195–216). Lincolnwood IL: Scientific Software International.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629. http://afhayes.com/spss-sas-and-r-macros-and-code.html

Hoyt, C. L. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6*, 153–160.

Kelley, K., & Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: evaluation of methods, recommendations, and software for composite measures. *Psychological Methods*, *21*(1), 69–92. https://doi.org/10.1037/a0040086

McNeish, D. (2018). Thanks coefficient alpha, We'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Millsap, R. E., & Everson, H. (1991). Confirmatory measurement model comparisons using latent means. *Multivariate Behavioral Research*, *26*(3), 479–497. https://doi.org/10.1207/s15327906mbr2603_6

O'Grady, K. E., & Medoff, D. R. (1991). Rater reliability: a maximum likelihood confirmatory factor-analytic approach. *Multivariate Behavioral Research*, *26*(2), 363–387. https://doi.org/10.1207/s15327906mbr2602_9

Peters, G. J. Y. (2014). The alpha and the omega of scale reliability and validity: why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *European Health Psychologist*, *16*(2), 56–69. https://doi.org/10.31234/osf.io/h47fv Supplemental materials and data available at https://osf.io/tnrxv/.

Raykov, T. (2007). Reliability if deleted, not 'alpha if deleted': evaluation of scale reliability following component deletion. *British Journal of Mathematical and Statistical Psychology*, *60*, 201–216. https://doi.org/10.1348/000711006X115954

Raykov, T. (2008). Alpha if item deleted: a note on loss of criterion validity in scale development if maximizing coefficient alpha. *British Journal of Mathematical and Statistical Psychology*, *61*, 275–285. https://doi.org/10.1348/000711007X188520

Shaw, B. P. (2021). Meeting assumptions in the estimation of reliability. *The Stata Journal*, *21*(4), 1021–1027.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37. https://doi.org/10.1037/1082-989X.8.1.16

Computational Guide 2

Item Response Theory

CHAPTER MENU

1 Item Response Theory (IRT) Models for Dichotomous Items, 400
2 Logistic (Logit) Item Response Theory Models, 400
  2.1 The Rasch Model, 401
  2.2 The One-Parameter Logistic (1PL) Model, 401
  2.3 The Two-Parameter Logistic (2PL) Model, 402
3 Data Example, 403
  3.1 Analyzing the Compressed Data, 403
  3.2 Expanding the Compact Data and Analyzing the Expanded Data, 404
  3.3 Estimating the Rasch Model, 405
  3.4 Item Response Theory (IRT) and Categorical Item Factor Analysis (CIFA) for Dichotomous Items (1PL), 406
4 Categorical Item Factor Analysis and Item Response Theory Models for Dichotomous Items, 407
  4.1 The Rasch Model, 410
  4.2 Estimated Ability Scores with the Rasch Model, 411
  4.3 Relative Model Fit of the 1PL and 2PL Models, 412
  4.4 Item Characteristic Curves (ICCs) and Item Information Functions (IIFs), 412
5 Normal Ogive (Probit) IRT Models, 414
6 IRT Models for Dichotomous Items Using Mplus, 415
  6.1 The Rasch Model, 415
  6.2 The 1PL Model, 416
  6.3 The 2PL Model, 417
  6.4 IRT Parameterization for the 1PL and Logistic Rasch Models, 417
7 IRT Models for Ordered Categorical (Polytomous) Items, 417
  7.1 Category Comparisons, 417
  7.2 Model Specification, 418
  7.3 Categorical Item Factor Analysis (CIFA) for Ordered Polytomous Categorical Items, 418
  7.4 Model Equivalence, Nesting, and Comparisons, 419
8 Four Ordinal Polytomous IRT Models, 421
  8.1 Model A: Item Discrimination Parameters and Step Thresholds, 421
  8.2 Model B: Common Discrimination Parameter and Step Thresholds, 421
  8.3 Model C: Item Discrimination Parameters with Item Location and Common Threshold Offsets, 422
  8.4 Model D: Common Discrimination Parameter with Item Location and Common Threshold Offsets, 422
9 Software and Example Analyses, 422
10 Cumulative Probability Models Using Stata, 423
  10.1 Model A: The Graded Response Model (GRM), 423
  10.2 Model B: The Graded Response Model with Common Discrimination (GRM-C), 424
11 Adjacent Category Models Using Stata, 424
  11.1 Model A: The Generalized Partial Credit Model (GPCM), 424
  11.2 Model B: The Partial Credit Model (PCM), 425

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

11.3 Model C: The Generalized Rating Scale Model (GRSM), 425
11.4 Model D: The Rating Scale Model (RSM), 425
11.5 Summary, 425
12 Cumulative Probability Models Using Mplus, 426
12.1 Model A: The Graded Response Model (GRM), 426
12.2 Model B: The Graded Response Model with Common Discrimination, 427
12.3 Model C: The Modified Graded Response Model (MGRM), 427
12.4 Model D: The Modified Graded Response Model with Common Threshold Offsets (MGRM-C), 428
13 Adjacent Category Models with Mplus, 429
13.1 Model A: The Generalized Partial Credit Model (GPCM), 429
13.2 Model B: The Partial Credit Model (PCM), 430
13.3 Model C: The Generalized Rating Scale Model (GRSM), 431
13.4 Model D, 431
14 Model Implementation, 432
14.1 Cumulative Probability Models, 432
14.2 Adjacent Category Models, 432
15 Model Fit for the Example Data, 433
A. Appendix, 434
References, 435

The Item Response Theory (IRT) methods covered in this computational guide are based fundamentally on the probability of responses to categorical items that are treated as indicators of an underlying latent trait. Like latent variable true score models, latent trait IRT models employ responses to multiple indicators to develop, refine, and evaluate measures of theoretical constructs or other attributes of individuals. We first introduce models for dichotomous or binary items. These were not only developed most intensively in educational measurement but are also used in psychology, social sciences, health, management, and other fields. We describe the differences and similarities between such well-known IRT models as the Rasch model, the one-parameter logistic (1PL) model, and the two-parameter logistic (2PL) model, along with their counterpart normal ogive models. Three- and four-parameter IRT models for binary items are beyond the scope of this introductory treatment. We also describe and apply familiar models for ordered categorical or polytomous data, including graded response models (GRMs) and partial credit models (PCMs).

1 Item Response Theory (IRT) Models for Dichotomous Items

In IRT with binary items, the probability of answering an item correctly is modeled as a simple or scaled discrepancy between the person's ability and the difficulty of the item. Where the observed data fit the model, individuals with higher ability are more likely to be able to answer difficult items correctly. The difficulty is the point on the ability continuum where the probability of correctly answering the item is 0.50. The relationship between ability and the probability of a correct answer is described by the Item Characteristic Curve or ICC. Figure 2.1 contains ICCs for five binary items that differ in their difficulty, but have a common discrimination parameter.

Specifically, the difficulty of an item is defined as the point on the latent trait (logit Theta) scale at which the probability of answering the question correctly is 0.50. In this case, items 1, 2, 3, 4, and 5 have difficulties of -2, -1, 0, 1, and 2, respectively.

2 Logistic (Logit) Item Response Theory Models

If an item is scored 1 when it is answered correctly and 0 when answered incorrectly, it is convenient to model the expected probabilities with the logit transformation by taking the logarithm of the probability of a correct answer divided by one minus the probability.

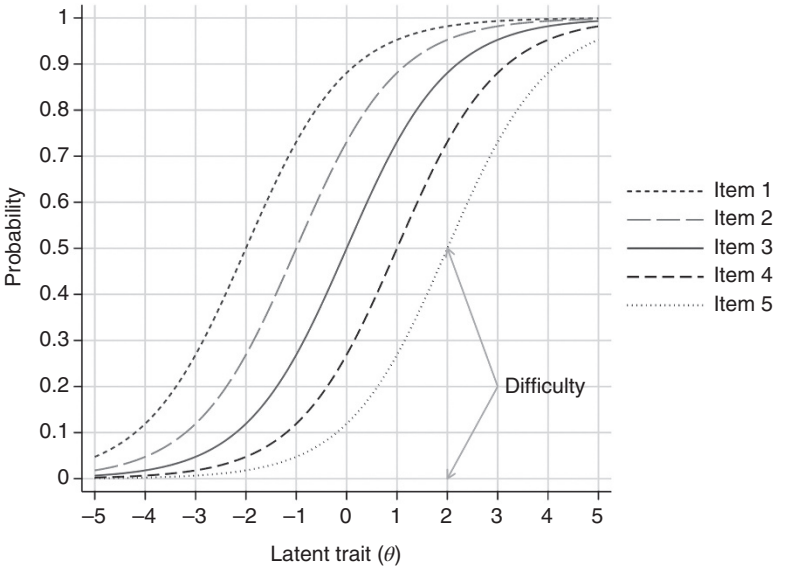


Figure 2.1 Logistic item characteristic curves for five (equally discriminating) items differing only in their difficulty (1PL model).

$$logit = ln\left[\frac{P(x=1)}{P(x=0)}\right] = ln\left[\frac{P(x=1)}{1-P(x=1)}\right]$$

A number of different logistic IRT models have been developed for dichotomous items. We will consider the Rasch model, the 1PL model, and the 2PL model, noting that other more complex or elaborate models, such as the three (3PL) and four (4PL) parameter logistic models, are available to address special measurement situations.

2.1 The Rasch Model

In the Rasch model, the logit of a dichotomous item j is expressed in terms of a traditional IRT linear predictor $\theta - \beta_j$ representing the difference between a person's latent trait or ability θ and item j's estimated difficulty β_j. The Rasch model also fixes the mean of the latent trait distribution to zero, $\mu_\theta = 0$, and freely estimates the latent trait variance σ^2_θ as a model parameter.

$$ln\left[\frac{P(x=1)}{1-P(x=1)}\right] = \theta - \beta_j$$

2.2 The One-Parameter Logistic (1PL) Model

The 1PL model is expressed in terms of the IRT linear predictor $\alpha(\theta - \beta_j)$. The Rasch model is equivalent in fit, prediction, and data likelihood to a 1PL model having a common discrimination parameter α, where the latent trait variance is fixed to one, $\sigma^2_\theta = 1$, and the latent trait mean is fixed to zero, $\mu_\theta = 0$. In this case, α may also be thought of as an estimated parameter for multiplicatively scaling the difference between the individual's latent trait and the item's difficulty. The estimated latent trait variance from the Rasch model is equal to the squared discrimination parameter from the 1PL model. ${}_{Rasch}\sigma^2_\theta = {}_{1PL}\alpha^2$.

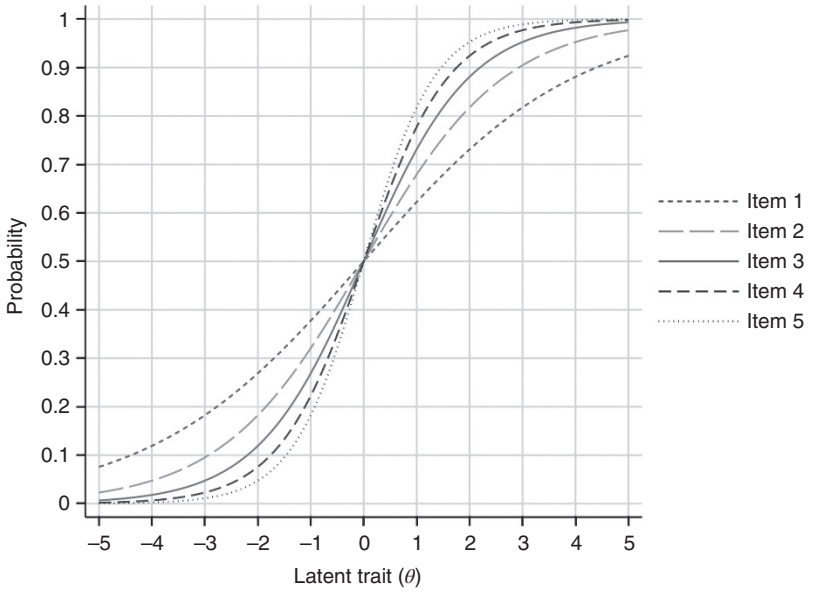


Figure 2.2 Items with the same difficulty, but differing in their discrimination parameters (2PL model).

$$ln\left[\frac{P(x=1)}{1-P(x=1)}\right] = \alpha\left(\theta - \beta_j\right)$$

2.3 The Two-Parameter Logistic (2PL) Model

The 2PL model generalizes the 1PL model by allowing separate discrimination parameters α_j for each item using the IRT linear predictor $\alpha_j(\theta - \beta_j)$. Like the 1PL model, the latent trait variance in the 2PL model is fixed to 1, $\sigma_\theta^2 = 1$, and the latent trait mean is fixed to zero, $\mu_\theta = 0$. In the 2PL model, items can vary in their discrimination parameters as well as their difficulties.

$$ln\left[\frac{P(x=1)}{1-P(x=1)}\right] = \alpha_j\left(\theta - \beta_j\right)$$

Items with higher discrimination α_j have a "steeper" slope while those with lower discrimination have a more gradual slope. Items with higher discrimination are better able to distinguish between those with differing latent trait levels. In Figure 2.2, all of the items have the same difficulty (0 on the logit Theta scale). One way to think about the discrimination of an item is to compare the probability of a correct response for those who have a Theta value at a fixed distance below and above the item's difficulty. For example, the probability of a correct response on item 1 for those with an ability of −1 on the logit Theta scale is about 0.38, while the probability of a correct response for those at a 1 on the scale is about 0.62. By comparison, with item 5, the probability of a correct response for those at −1 on the logit Theta scale is about 0.18, whereas the probability of a correct response for those at 1 on the scale is about 0.82. Item 5 is thus better able than item 1 to discriminate or distinguish between people of similar but differing ability levels.

3 Data Example

IRT procedures in Stata and SAS can be used to estimate the parameters of the 1PL and 2PL models, along with dedicated IRT programs such as IRTPRO, flexMIRT, and generalized latent variable modeling programs such as Mplus or Stata's *gsem*. We will use a classic dataset with five binary items from section 6 of the Law School Admission Test (Bock & Lieberman, 1970, p. 188) to illustrate the application of IRT for dichotomous items. The original Bock & Lieberman data are given in the form of cell frequencies for each of the 32 response profiles resulting from jointly tabulating the five binary items. IRT software that can apply frequency weights, such as Stata irt, can directly analyze this data. The data in their original form are given in the appendix to this guide. To prepare the data for analysis in Stata, it is convenient to import the space-delimited data table and recode the original responses of 1 and 2 to the more usual representation of 0 and 1, where 1 in the new format represents a correct response.

```
import delimited lsat6.dat, /// Input data file
 delimiter (" ") clear      //  Space delimited
recode v1-v5 (1=0) (2=1)    //  Rescale from 1/2 to 0/1
```

3.1 Analyzing the Compressed Data

Where the five binary items defining the response profiles are in variables v1 through v5, and v6 contains the frequency counts for each response profile, the following Stata irt command can be used to estimate the 1PL model for the original frequency count data [IRTBinary.do].

```
*** 1PL model via Stata irt using compressed data and frequency
weights
irt 1pl v1-v5 [fweight=v6]
```

The *fweight* specification in square brackets on the *irt 1pl* command indicates that variable v6 contains the frequency counts for the 32 different response profiles, giving a total sample size of 1000 as the total number of individuals being equal to the sum of these weights. The associated output gives the item difficulties and the common discrimination parameter. The discrimination parameter was 0.755 with difficulties ranging from −3.615 to −0.318.

	One-parameter logistic model log-likelihood = −2466.9376				Number of obs = 1,000		
		Coef.	Std. Err.	z	P>|z|	[95% Conf.	interval]
	Discrim	0.7551283	0.0694206	10.88	0.000	0.6190666	0.8911901
v1	Diff	−3.615293	0.3265991	−11.07	0.000	−4.255416	−2.975171
v2	Diff	−1.322434	0.1421673	−9.30	0.000	−1.601077	−1.043792
v3	Diff	−0.3176353	0.0976766	−3.25	0.001	−0.5090779	−0.1261928
v4	Diff	−1.730106	0.1691149	−10.23	0.000	−2.061565	−1.398647
v5	Diff	−2.780193	0.251015	−11.08	0.000	−3.272174	−2.288213

Using the compressed data layout described above, the first three of the 32 rows of data containing information for the 11 respondents having the first 3 response patterns are shown here. Three of the respondents answered all of the items incorrectly, another six respondents answered all of the

items incorrectly except number five, and two respondents answered all of the items incorrectly except item 4.

v1	v2	v3	v4	v5	v6
0	0	0	0	0	3
0	0	0	0	1	6
0	0	0	1	0	2
...					

This data set is compact with just 32 rows or observations but requires weighted analysis using the frequency weights in variable v6. Because two of the original 32 rows have weights of 0, they are not actually used in the analysis. In many cases, it can be convenient to expand the data from this collapsed form so that each row or record in the expanded dataset represents one of the 1000 individuals in this data, each having the associated five item responses.

3.2 Expanding the Compact Data and Analyzing the Expanded Data

With Stata, the data can be expanded as shown here. We first *drop* any cases with a zero frequency count in v6, *expand* the data using the counts, and then *drop* variable v6. For convenience, we sort the cases by their response profiles and generate an ID variable for the 1000 individual respondents. We also create a total number correct score for later use [IRTBinary.do].

```
*** Expand compressed data in Stata
drop if v6==0  // Drop empty cells
expand v6      // Expand the data
drop v6        // Drop the frequency variable
sort v1-v5     // Recover the original data order
generate id=_n // Generate a person identifier
egen tot=rowtotal(v1-v5) // Number correct score
```

The resulting expanded response data for the same 11 respondents is shown below. This dataset has 1000 rows or observations but can be analyzed without using the frequency weights.

id	v1	v2	v3	v4	v5
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	1
5	0	0	0	0	1
6	0	0	0	0	1
7	0	0	0	0	1
8	0	0	0	0	1
9	0	0	0	0	1
10	0	0	0	1	0
11	0	0	0	1	0
	...				

Although the person identifier is arbitrary based on the ordering of the data, persons with id numbers 1 through 3 missed all of the items, persons 4 through 9 missed all but item 5, and persons 10 and 11 missed all but item 4. With a small number of items, the collapsed data is more compact but requires weighted analysis, whereas the expanded data is more voluminous but does not require weighting in the analysis. However, as the number of items increases relative to the sample size, the expanded or person-level data will generally be more suitable.

The 1PL model can be estimated with the expanded data, giving results that are identical to those for the weighted analysis of the compressed data [IRTBinary.do].

```
*** 1PL model for the expanded data using Stata irt
irt 1pl v1-v5 // 1PL model IRT parameters
```

It is also useful to note that the 1PL model can also be obtained by equating the discrimination parameters of the 2PL model using symbolic parameter constraints in Stata irt. This is accomplished by using the *cns(a@a)* option on the *irt 2pl* command. This specifies that the discrimination parameter (the a before the @ sign) is set to a single value for all of the items (the a after the @ sign.

```
*** 1PL model as a constrained 2PL model with a common discrimination
parameter
irt 2pl v1-v5, cns(a@a) // 1PL model as a constrained 2PL model
```

3.3 Estimating the Rasch Model

Although Stata's irt procedure does not directly estimate the Rasch model, Stata's *gsem* can be used instead to obtain the Rasch model IRT parameters, with the exception that the estimated intercepts are the negative of the standard IRT difficulty parameters. In this *gsem* command, the list of items (v1–v5) is followed with an @ sign and the number 1. This constrains the discrimination parameter to be equal to 1.0 for all of the items. Note that the *gsem* command uses the *logit* option to indicate that binary items are being analyzed.

```
 *** Estimate the Rasch model via Stata's gsem command
gsem (F -> v1-v5@1), logit     // Rasch model CIFA parameters
```

The resulting output gives the (negative of the) estimated item difficulties for each item, along with the estimated latent trait variance $\sigma^2_\theta = 0.570$. As mentioned earlier the estimated latent trait variance from the Rasch model is equal to the squared discrimination parameter from the 1PL model, ${}_{Rasch}\sigma^2_\theta = {}_{1PL}\alpha^2$, or in this case, $0.570 = 0.755^2$.

	Coef.	Std. Err.	z	P>|z|	[95% Conf. interval]	
v1	12.730011	0.1304396	20.93	0.000	2.474354	2.985667
v2	1.9986077	0.0791776	12.61	0.000	0.8434225	1.153793
v3	1.2398554	0.0717749	3.34	0.001	0.0991792	0.3805316
v4	11.306452	0.0846381	15.44	0.000	1.140565	1.47234
v5	12.099403	0.1054439	19.91	0.000	1.892737	2.306069
var(F)	0.5702189	0.1048429			0.3976821	0.8176117

The traditional IRT difficulty estimates and their standard errors, significance tests and confidence intervals can be obtained by using Stata's *nlcom* command for each item.

```
 *** Obtain the Rasch model IRT parameters via Stata's gsem command
forvalues j=1/5 {              // Get the Rasch IRT difficulty
estimates
 nlcom (b`j': -_b[v`j':_cons]/sqrt(_b[/var(F)]))
}
```

3.4 Item Response Theory (IRT) and Categorical Item Factor Analysis (CIFA) for Dichotomous Items (1PL)

The standard 1PL IRT model uses the linear predictor $\alpha(\theta - \beta_j)$ to model the logit transformation of the response probability for dichotomous items. However, Stata's irt estimation procedure instead internally uses the linear predictor $\alpha\theta + \beta_j^*$ where $\beta_j^* = -\beta_j \bullet \alpha$. This is used to conduct a CIFA, and then the CIFA parameters are transformed into standard IRT form to display them in the Stata output. For example, we can convert the 1PL internal CIFA difficulty parameter for v2 into the corresponding IRT parameter $\beta_2 = -\beta_2^*/\alpha = -0.999/0.755 = 1.322$. Note that the IRT estimate α of the common discrimination parameter for the 1PL model is the same as the CIFA estimate.

	1PL Parameters		
	IRT	CIFA	
α	0.755	0.755	α
β_1	−3.615	2.730	β_1^*
β_2	−1.322	0.999	β_2^*
β_3	−0.318	0.240	β_3^*
β_4	−1.730	1.306	β_4^*
β_5	−2.780	2.099	β_5^*

The CIFA parameters can be obtained by using the following Stata *irt 1pl* command with the *estmetric* option [IRTBinary.do]:

```
*** Categorical item factor analysis (CIFA) parameters for the 1PL model
irt 1pl v1-v5, estmetric // CIFA parameters
```

Alternatively, Stata's *gsem* command can be used to obtain the same CIFA parameter estimates. The@ sign followed by the letter a after the list of variables indicates that a common discrimination parameter should be estimated. Notice also that the variance of the latent trait is fixed to 1.0 with the option *variance(F@1)*.

```
*** Obtain the 1PL CIFA parameters via Stata's gsem command
gsem (F -> v1-v5@a), variance(F@1) logit
```

Only the CIFA parameters (and not the IRT parameters) are available for subsequent use in post-estimation computations following Stata's irt command. The internal labels for the CIFA parameters can be obtained by using the following Stata *irt 1pl* command with the *coeflegend* option:

```
*** Get the internal CIFA parameter labels
irt 1pl v1-v5, coeflegend
```

Here, the CIFA discrimination parameter has the same common value (0.755) for all of the items, but separate item difficulty parameters are estimated for each item.

		Coef.	Legend
v1	Theta	0.7551283	_b[v1:Theta]
	_cons	2.730011	_b[v1:_cons]
v2	Theta	0.7551283	_b[v2:Theta]
	_cons	0.9986076	_b[v2:_cons]
v3	Theta	0.7551283	_b[v3:Theta]
	_cons	0.2398554	_b[v3:_cons]
v4	Theta	0.7551283	_b[v4:Theta]
	_cons	1.306452	_b[v4:_cons]
v5	Theta	0.7551283	_b[v5:Theta]
	_cons	2.099403	_b[v5:_cons]
	var(Theta)	1	_b[/var(Theta)]

One can also directly obtain the CIFA parameters for the 1PL model by using Stata's *gsem* command. The model specification here equates the factor loadings (a) and fixes the variance of the latent trait to one.

```
*** Obtain CIFA parameters via Stata's gsem command
gsem (F -> v1-v5@a), variance(F@1) logit
```

4 Categorical Item Factor Analysis and Item Response Theory Models for Dichotomous Items

It is possible to conceptualize and estimate IRT models as transformed CIFA models. As before, the traditional 2PL IRT model can be expressed as

$$ln\left[\frac{P(x=1)}{1-P(x=1)}\right] = \alpha_j\left(\theta - \beta_j\right)$$

The corresponding logit CIFA model can be expressed as

$$ln\left[\frac{P(x=1)}{1-P(x=1)}\right] = \alpha_j\theta + \beta_j^*$$

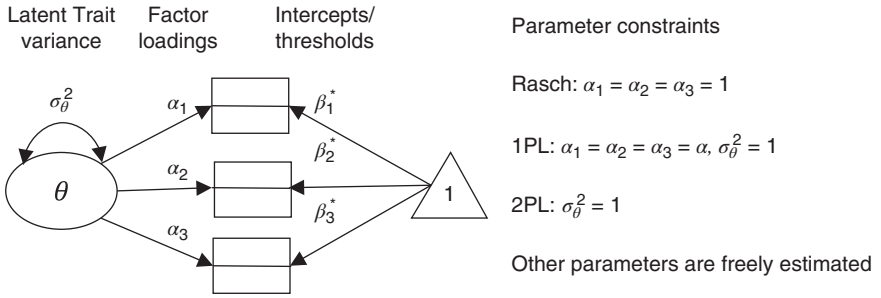


Figure 2.3 Using CIFA to estimate selected IRT models for dichotomous items.

Here θ is the latent trait, α_j are item-specific factor loadings, and β_j^* are the item-specific threshold values. The CIFA factor loadings are equal to the IRT item discrimination parameters and the CIFA item thresholds are related to the IRT item difficulties as follows:

$$\beta_j = -\beta_j^*/\alpha_j$$

Figure 2.3 depicts the use of CIFA as a modeling framework for dichotomous items. In this path diagram, an oval is used to indicate the latent trait and divided rectangles are used to indicate the dichotomous items. The item discrimination parameters α_j are indicated as factor loadings accompanying the straight single-headed arrows linking the latent trait to the items. The CIFA item intercept parameters β_j^* are indicated as thresholds accompanying the straight single-headed arrows linking the triangle containing a 1 with the items. Finally, the latent trait variance σ_θ^2 accompanies the curved double-headed arrow with both ends connected to the latent trait.

The Rasch model is obtained by constraining the factor loadings to 1 and freely estimating the latent trait variance σ_θ^2, the 1PL model is obtained by equating the factor loadings and fixing the latent trait variance to 1, and the 2PL model is obtained by fixing the latent trait variance to 1 but estimating item-specific loadings α_1, α_2, and α_3. The thresholds β_1^*, β_2^*, and β_3^* are freely estimated in all three models.

CIFA and IRT parameterization. Stata's *irt* actually performs its estimation using the CIFA specification and then transforms the results from CIFA to IRT parameterization (see Table 2.1). For example, the following *irt* commands give the traditional IRT difficulty and discrimination and the CIFA intercept and slope parameterization for the 1PL and 2PL models.

```
*** IRT difficulty-discrimination and CIFA intercept-slope
parameterization
irt 1pl v1-v5  // IRT 1PL Difficulty-discrimination parameterization
irt 2pl v1-v5  // IRT 2PL Difficulty-discrimination parameterization
irt 1pl v1-v5, estmetric // CIFA 1PL Intercept-slope parameterization
irt 2pl v1-v5, estmetric // CIFA 2PL Intercept-slope parameterization
```

The IRT and CIFA specifications of the 2PL model are alternate parameterizations of the same statistical model and necessarily have the same log-likelihood. Stata *irt* automatically transforms the original CIFA parameters into the corresponding IRT parameters.

Table 2.1 CIFA and IRT intercept/difficulty and slope/discrimination parameters for the 1PL and 2PL models using Stata irt.

	CIFA (estmetric)				IRT			
	1PL		2PL		1PL		2PL	
	est	se	est	se	est	se	est	se
Intercept/difficulty								
v1	2.730	0.130	2.773	0.206	−3.615	0.327	−3.359	0.867
v2	0.999	0.079	0.990	0.090	−1.322	0.142	−1.370	0.307
v3	0.240	0.072	0.249	0.076	−0.318	0.098	−0.280	0.100
v4	1.306	0.085	1.285	0.099	−1.730	0.169	−1.866	0.434
v5	2.099	0.105	2.053	0.135	−2.780	0.251	−3.126	0.871
Slope/discrimination								
v1	0.755	0.069	0.826	0.258	0.755	0.069	0.826	0.258
v2	0.755	0.069	0.723	0.187	0.755	0.069	0.723	0.187
v3	0.755	0.069	0.891	0.233	0.755	0.069	0.891	0.233
v4	0.755	0.069	0.688	0.185	0.755	0.069	0.688	0.185
v5	0.755	0.069	0.657	0.210	0.755	0.069	0.657	0.210
Latent variance	1.000	—	1.000	—	1.000	—	1.000	—

The CIFA equivalent of the 2PL model can also be estimated using the Stata generalized structural equation modeling (*gsem*) procedure. The *gsem* command specifies a single latent variable, F, that is measured by the five items using the logit link and fixing the latent trait variance to 1. The CIFA factor loadings are equal to the corresponding IRT discrimination parameters. Stata's *nlcom* command is then used to calculate the IRT item difficulty estimates from the corresponding CIFA threshold and factor loading estimates. We have used Stata's *forvalues* structure to illustrate how this approach could be extended to a potentially large number of items.

```
*** 2PL model as a CIFA using Stata gsem
gsem(F->v1-v5), logit variance(F@1)    // CIFA 2PL equivalent model

forvalues j=1/5 {                      // Get the 2PL IRT difficulties
                                          from
nlcom (b`j': -_b[v`j':_cons]/_b[v`j':F]) // the CIFA thresholds & loadings
}
```

The parameters of the 1PL model can similarly be obtained using *gsem* by equating the factor loadings and fixing the factor variance to 1. The parameters of the Rasch model can be estimated using *gsem* by fixing the factor loadings to 1 but by freely estimating the factor variance (by omitting the constraint fixing the variance to 1).

```
*** Rasch model via CIFA using Stata gsem
gsem(F->v1-v5@1), logit        // CIFA Rasch equivalent model

forvalues j=1/5 {              // Get the Rasch IRT difficulties
 nlcom (b`j': -_b[v`j':_cons]) //  from the CIFA thresholds &
loadings
}
```

In Stata, the difficulty and discrimination parameters for the Rasch, 1PL, and 2PL models can be obtained using either the *irt* or *gsem* procedures. The *irt* procedure can be followed with *irtgraph* to produce item and test characteristic curve and information function plots, but these plots are not available with *gsem*.

4.1 The Rasch Model

Although Stata irt does not directly estimate the Rasch model, the Rasch and 1PL models are statistically equivalent, and we can potentially transform the parameters of the 1PL model into those for the Rasch model. By utilizing the CIFA parameters from the 1PL model, it is possible to obtain the parameters of the IRT Rasch model by using Stata's *nlcom* post-estimation command for estimating nonlinear functions and combinations of the estimated model parameters. The latent trait variance for the Rasch model is obtained by squaring the common 1PL discrimination parameter. The difficulty parameters displayed in the 1PL irt procedure output are appropriately transformed into the standard IRT 1PL parameters, but the internal CIFA difficulty parameters for the 1PL model have the opposite sign of the corresponding IRT Rasch model parameters.

```
*** Rasch model via Stata irt 1pl using wide form data
irt 1pl v1-v5                  // Estimate the 1PL model

 nlcom _b[v1:Theta]^2          // Get the Rasch latent trait variance
forvalues j=1/5 {              // Get the Rasch difficulty estimates
 nlcom (b`j': -_b[v`j':_cons])//  as the negative of 1PL intercepts
}
```

Stata irt's internal names for the discrimination and difficulty parameter for the first item are _b[v1:Theta] and _b[v1:_cons], respectively. The first *nlcom* command is used to square the discrimination parameter to produce the estimated variance of the latent trait, Theta. The *foreach* command structure is used to obtain the Rasch difficulty parameters as the negative of the internal difficulty parameters for each of the five items from the 1PL model, again using the *nlcom* command. Table 2.2 contains results for the Rasch, 1PL, and 2PL models using Stata irt.

The Rasch and 1PL models are statistically equivalent and represent alternate parameterizations of the same underlying statistical model. Note that the estimated latent trait variance from the Rasch model (0.570) is equal to the squared common discrimination parameter (0.755^2) from the 1PL model. These two models thus fit the data identically.

Table 2.2 IRT difficulty and discrimination parameters for the Rasch, 1PL, and 2PL models.

	Rasch		1PL		2PL	
	est	**se**	**est**	**se**	**est**	**se**
Difficulty						
v1	−2.730	0.130	−3.615	0.327	−3.359	0.867
v2	−0.999	0.079	−1.322	0.142	−1.370	0.307
v3	−0.240	0.072	−0.318	0.098	−0.280	0.100
v4	−1.306	0.085	−1.730	0.169	−1.866	0.434
v5	−2.099	0.105	−2.780	0.251	−3.126	0.871
Discrimination						
v1	1.000	—	0.755	0.069	0.826	0.258
v2	1.000	—	0.755	0.069	0.723	0.187
v3	1.000	—	0.755	0.069	0.891	0.233
v4	1.000	—	0.755	0.069	0.688	0.185
v5	1.000	—	0.755	0.069	0.657	0.210
Latent variance	0.570	0.105	1.000	—	1.000	—

4.2 Estimated Ability Scores with the Rasch Model

One of the central goals of many IRT applications is to provide estimates of the underlying latent trait or ability of each person. In Stata irt, the estimated or predicted latent trait values $\hat{\theta}$ can be obtained with the following syntax.

```
*** Obtain the Rasch Model ability (theta) estimates
irt 1pl v1-v5                    // Estimate the 1PL model
predict onepl, latent            // Estimated 1PL latent trait values
generate r1=onepl*_b[v1:Theta] // Rescale 1PL to Rasch abilities
egen tot=rowtotal(v1-v5)         // Calculate number correct total
score
tab r1 tot                       // Display Rasch ability estimates
correlate r1 tot                 // Correlation between total score &
theta hat
```

The *predict* command is used to obtain the predicted or estimated latent trait values from the 1PL model, and then the *generate* command transforms these estimates to the Rasch scale. One frequently cited advantage of the Rasch model is that when the data fit the model, the total score or number correct score is a sufficient statistic for the predicted latent trait values. One way to explore this is to examine the correspondence between the estimated latent trait scores and the total or number correct scores. The rows represent the estimated Rasch model latent trait scores while the columns give the total scores. For each observed total score, there is a single estimated latent trait value. The estimated correlation between the ability estimates and the total score is 0.9997.

	tot						
r1	0	1	2	3	4	5	Total
−1.442382	3	0	0	0	0	0	3
−1.078932	0	20	0	0	0	0	20
−0.7102423	0	0	85	0	0	0	85
−0.3312488	0	0	0	237	0	0	237
0.0630738	0	0	0	0	357	0	357
0.4773941	0	0	0	0	0	298	298
Total	3	20	85	237	357	298	1,000

4.3 Relative Model Fit of the 1PL and 2PL Models

The 1PL and 2PL models can be compared in terms of a Chi-square difference test, and information criteria such as AIC or BIC. In Stata irt, the following syntax can be used to make these comparisons. The `estat ic` command is used to obtain the information criteria (AIC & BIC) for each model, and the `estimates store` command stores the estimates for each model in order to compute a likelihood ratio test comparing the two models using the `lrtest` command.

```
*** Model comparison test for 1PL and 2PL models
irt 1pl v1-v5         // One-parameter Logistic IRT model
estat ic              // Get information criteria
estimates store onepl // Store the 1pl estimates
irt 2pl v1-v5         // Two-parameter Logistic IRT model
estat ic              // Get information criteria
estimates store twopl // Store the 2pl estimates
lrtest onepl twopl    // 1PL vs. 2PL Chi-square difference test
```

The likelihood ratio Chi-square difference test confirms that the more parsimonious 1PL model fits the data just as well (see Table 2.3) as the 2PL model ($\chi^2_{diff}(4) = 0.057, p = 0.966$).

4.4 Item Characteristic Curves (ICCs) and Item Information Functions (IIFs)

In addition to fitting the models and estimating the model parameters, Stata irt can provide the associated item and test characteristic curve and information function plots. A plot of the item characteristic curves (ICCs) for the five items (see Figure 2.4) can be easily obtained with the irtgraph command following the estimation of the 1PL model in Stata.

Table 2.3 Information criteria for the 1PL and 2PL models for the LSAT6 data.

	−2 Log-likelihood	Number of parameters	AIC	BIC
1PL	−2466.938	6	4945.875	4975.322
2PL	−2466.654	10	4953.307	5002.385

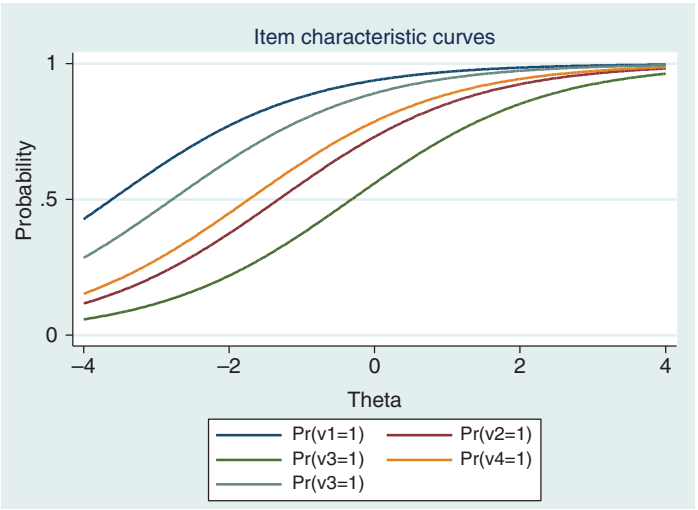


Figure 2.4 Item characteristic curves for the 1PL model for the five LSAT6 items using Stata irtgraph.

```
irt 1pl v1-v5          // One-parameter Logistic IRT model
irtgraph icc
```

The item information curves for the five LSAT6 items (see Figure 2.5) can be obtained with the following command.

```
irtgraph iif
```

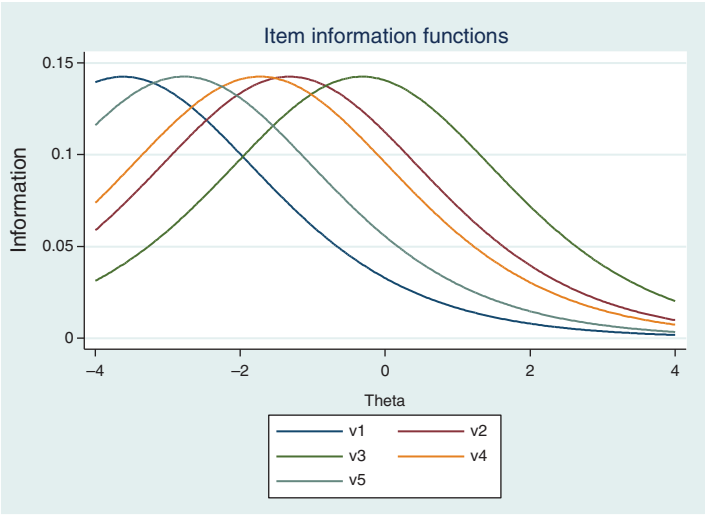


Figure 2.5 Item information function curves for the 1PL model for the five LSAT6 items using Stata irtgraph.

5 Normal Ogive (Probit) IRT Models

An alternative to using a logistic IRT model is to use a cumulative normal probability IRT model. The 2PL model was expressed earlier in terms of a logit transformation of the expected probability of answering an item correctly.

$$ln\left[\frac{P(x=1)}{1-P(x=1)}\right] = \alpha_j\left(\theta-\beta_j\right)$$

This can be equivalently expressed in terms of the logistic distribution function as

$$P(x=1 \mid \theta) = \frac{\exp\left[\alpha_j\left(\theta-\beta_j\right)\right]}{1+\exp\left[\alpha_j\left(\theta-\beta_j\right)\right]} = \boldsymbol{\Psi}\left[\alpha_j\left(\theta-\beta_j\right)\right]$$

where $\boldsymbol{\Psi}$ is the logistic distribution function expressing the expected probability of a correct item response in terms of the traditional item difficulty and discrimination parameters. Instead of using the logistic distribution function, one can use the cumulative normal probability or probit function $\boldsymbol{\Phi}$ to relate the probability to the difficulty and discrimination parameters.

$$(x=1 \mid \theta) = \boldsymbol{\Phi}\left[\alpha_j^{**}\left(\theta-\beta_j^{**}\right)\right]$$

Double asterisk superscripts (**) are used to distinguish probit from logit discrimination and difficulty parameters. Such models are often called normal ogive or probit IRT models, and provide results very similar to logistic IRT models, with the exception of a scaling factor.

$$\alpha_j \approx 1.702\left(\alpha_j^{**}\right)$$

$$\beta_j \approx 1.702\left(\beta_j^{**}\right)$$

The form of the normal distribution function is very similar to the standardized logistic distribution function (see Figure 2.6).

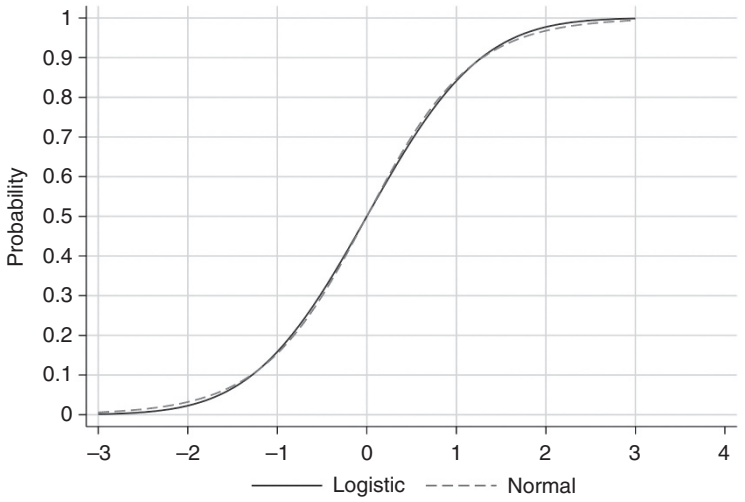


Figure 2.6 Relationship between the standardized logistic and normal ogive functions.

Although Stata's irt procedure does not estimate normal ogive (probit) IRT models, these are available by using *gsem* in Stata. For example, the two-parameter normal ogive (2PNO) model can be obtained with the following syntax. The only change from the 2PL model is to replace the *logit* option with the *probit* option on the *gsem* command.

```
*** Two parameter normal ogive (2PNO) model using Stata gsem
gsem(F->v1-v5), probit variance(F@1)     // CIFA 2PNO equivalent model

forvalues j=1/5 {                        // Get IRT difficulties from the
nlcom (b`j': -_b[v`j':_cons]/_b[v`j':F]) // CIFA thresholds & loadings
}
```

6 IRT Models for Dichotomous Items Using Mplus

6.1 The Rasch Model

In the Rasch model, the logit of a dichotomous item j is expressed in terms of a traditional IRT linear predictor $\theta - \beta_j$ representing the difference between a person's latent trait or ability θ and the item's estimated difficulty β_j.The Rasch model also fixes the latent trait mean to zero, $\mu_\theta = 0$, and estimates the latent trait variance σ^2_θ as a model parameter.

$$ln\left[\frac{P(x=1)}{P(x=0)}\right] = ln\left[\frac{P(x=1)}{1-P(x=1)}\right] = ln\left[\frac{P}{1-P}\right] = \theta - \beta_j$$

To simplify estimation, some IRT software programs transform the usual difficulty parameterization into an equivalent intercept parameterization with the linear predictor $\theta + \nu_j$, where $\nu_j = -\beta_j$. However, Mplus fixes the intercepts $\nu_j = 0$, and instead uses a direct threshold parameterization with the linear predictor $\theta + \tau_j$. The Rasch model can be estimated in Mplus as a constrained CIFA, treating θ as a latent variable and fixing all of the item factor loadings to 1.

```
TITLE: Binary item IRT models with frequency weights
DATA: FILE IS lsat6.dat;      ! Data file
VARIABLE: NAMES ARE v1-v6;    ! Variable names
 FREQWEIGHT IS V6;            ! Frequency weights
 USEOBSERVATIONS ARE (V6>0); ! Select positive weights
 USEVARIABLES ARE v1-v5;      ! Select variables
 CATEGORICAL ARE v1-v5;       ! Categorical variables
ANALYSIS: ESTIMATOR IS MLR;   ! MLR Estimation
MODEL: f BY v1-v5@1;          ! Rasch model
PLOT: TYPE IS PLOT3;          ! ICC and information plots
```

The plot command produces the ICCs (Figure 2.7) and item information curves (Figure 2.8).

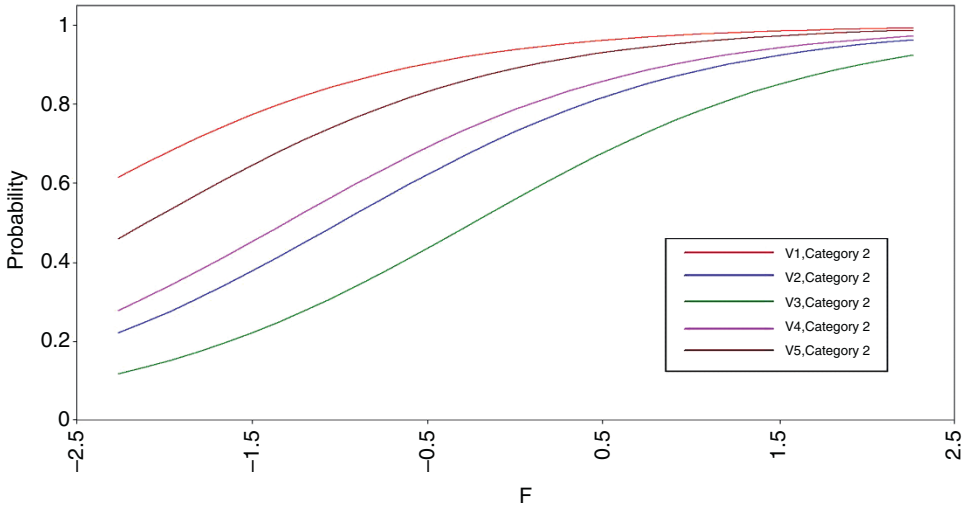


Figure 2.7 Mplus Rasch model item characteristic curves (ICCs) for the five Adapted from Bock & Lieberman (1970) items.

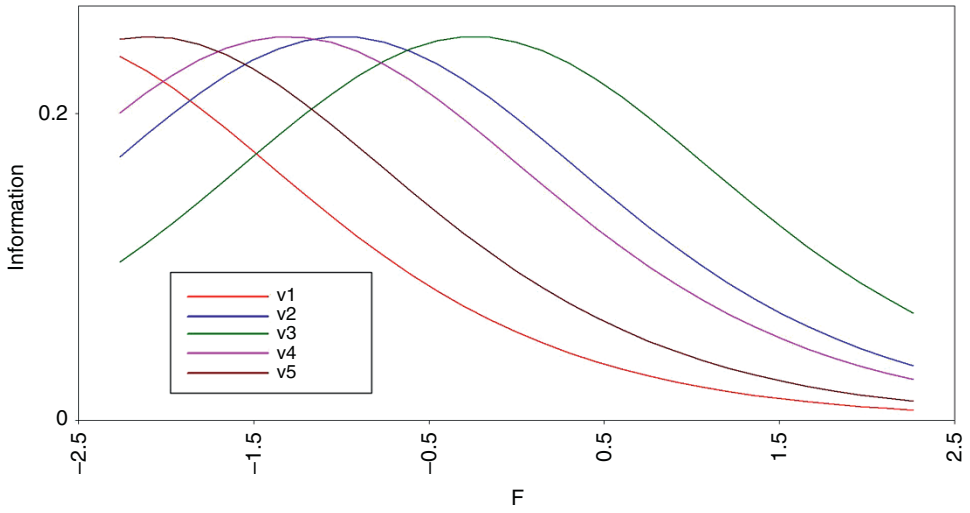


Figure 2.8 Mplus Rasch model item information curves (IICs) for the five Bock & Lieberman et al / (1970) / Springer Nature items.

6.2 The 1PL Model

The Rasch model is equivalent to a 1PL model having a common discrimination parameter α, where the latent trait variance is fixed to one $\sigma_\theta^2 = 1$. The 1PL IRT model uses the linear predictor $\alpha(\theta - \beta_j)$, and can be estimated in Mplus as a CIFA model with the linear predictor $\alpha\theta + \tau_j$ where $\tau_j = \beta_j/\alpha$. The Rasch and 1PL models are parameterized differently but are equivalent in fit, prediction, and data likelihood. The 1PL model can be specified as a CIFA model in Mplus by

estimating a common factor loading and fixing the variance of the latent trait to 1. This can be accomplished with the following modification to the MODEL: command in the previous program listing.

```
MODEL: f@1; f BY v1-v5* (a); ! 1PL model
```

The CIFA formulations of the Rasch and 1PL models estimate the same threshold parameters ${}_{Rasch}\tau_j = {}_{1PL}\tau_j$. The estimated latent trait variance from the Rasch model is equal to the squared discrimination parameter from the 1PL model ${}_{Rasch}\sigma^2_\theta = {}_{1PL}\alpha^2$. Equivalently, the discrimination parameter from the 1PL model is equal to the latent trait standard deviation from the Rasch model ${}_{1PL}\alpha = {}_{Rasch}\sigma_\theta$.

6.3 The 2PL Model

The 2PL model generalizes the 1PL model by allowing separate discrimination parameters α_j for each item using the IRT linear predictor $\alpha_j(\theta - \beta_j)$. The 2PL model can be estimated in Mplus as a CIFA model with item-specific factor loadings using the transformed linear predictor $\alpha_j\theta + \tau_j$, where $\tau_j = \beta_j/\alpha_j$. The traditional 2PL model can be specified by freely estimating all of the factor loadings and fixing the variance of the latent trait to 1. This can be done with the following modification to the Mplus MODEL: command.

```
MODEL: f@1; f BY v1-v5*;      ! 2PL model
```

For convenience, Mplus transforms the 2PL CIFA estimated parameters $\alpha_j\theta + \tau_j$ into the traditional IRT metric $\alpha_j(\theta - \beta_j)$, such that $\beta_j = \tau_j/\alpha_j$, and provides these as ancillary results in the output with the following heading.

```
IRT PARAMETERIZATION IN TWO-PARAMETER LOGISTIC METRIC
WHERE THE LOGIT IS DISCRIMINATION*(THETA - DIFFICULTY)
```

6.4 IRT Parameterization for the 1PL and Logistic Rasch Models

Mplus similarly provides estimated discrimination and difficulty parameters by transforming the estimated 1PL CIFA parameters $\alpha\theta + \tau_j$ into the traditional IRT metric $\alpha(\theta - \beta_j)$ such that $\beta_j = \tau_j/\alpha$. Mplus uses the latent trait standard deviation as a common discrimination parameter $\alpha^* = \sigma_\theta$ for the CIFA Rasch model to create transformed discrimination and difficulty parameters that are identical to those from the 1PL model such that $\beta_j = \tau_j/\alpha^*$.

7 IRT Models for Ordered Categorical (Polytomous) Items

7.1 Category Comparisons

Two main category comparison methods have been utilized in the development of IRT models for ordered polytomous items (Penfield, 2014; see Table 2.4)—cumulative probability comparisons and adjacent category comparisons.

Table 2.4 Category comparison methods for ordered polytomous items.

Category comparison methods		Steps	Categories			
Ordered polytomous items			1	2	3	4
Cumulative probability	$\frac{P(x \geq m)}{P(x < m)}$	1	1	2	3	4
		2	1	2	3	4
		3	1	2	3	4
Adjacent category	$\frac{P(x = m)}{P(x = m-1)}$	1	1	2		
		2		2	3	
		3			3	4

With M ordered categories, there are $K = M - 1$ possible comparisons of adjoining categories. Following Penfield (2014), we have elected to describe these comparisons using the concept of a shift or step from one category or combination of categories to the next, despite Andrich's (2015) concern about the theoretical appropriateness of applying the step concept to either cumulative probability or adjacent category comparisons. For each step comparison in Table 2.4, presence in the shaded categories is compared to presence in the unshaded ones. For example, using cumulative probability comparisons with four categories we would say that there are three steps, the first step contrasting the first category to the last three, the second step contrasting the first two with the last two, and the third step contrasting the first three categories with the last one. Similarly, with adjacent category comparisons, the first step is from the first category to the second one, the second step is from the second category to the third, and the third step is from the third category to the last one.

7.2 Model Specification

Eight common IRT models for polytomous items (see Table 2.5) can be described in terms of the combinations of the two main category comparison methods and four approaches to the structuring of the IRT discrimination and threshold parameters (Penfield, 2014; Bacci, Bartolucci, & Gnaldi, 2014). Cumulative probability models and adjacent category models are sometimes referred to respectively as difference models and divide-by-total models (Thissen & Steinberg, 1986), or as indirect and direct models (Embretson & Reise, 2000).

7.3 Categorical Item Factor Analysis (CIFA) for Ordered Polytomous Categorical Items

Models equivalent to IRT models A through D are frequently estimated using latent variable or factor analysis models for categorical items. CIFA models utilize what is often called the slope-intercept parameterization rather than the traditional IRT discrimination-threshold parameterization. When properly specified, the slope and intercept parameters of corresponding CIFA models (models A$*$ through D$*$) can generally be transformed into the traditional IRT discrimination and threshold parameters with appropriate software. In Table 2.6, some of the models implicitly impose equality constraints across items on the IRT discrimination or threshold parameters.

Table 2.5 Polytomous IRT models by category comparison methods and model parameter structure.

Model	Basic linear predictor	Constraints	Model parameters		Cumulative probability	Adjacent category
			Discrimination	Threshold		
A	$\alpha_j(\theta - \beta_{jk})$		Item	Item/Step	GRM	GPCM
B	$\alpha(\theta - \beta_{jk})$		Common	Item/Step	GRM-C	PCM
C	$\alpha_j(\theta - \delta_j - \tau_k)$	$\sum_{k=1}^{K}\tau_k = 0$	Item	Scale	MGRM	GRSM
D	$\alpha(\theta - \delta_j - \tau_k)$	$\sum_{k=1}^{K}\tau_k = 0$	Common	Scale	MGRM-C	RSM

Note 1 (Parameters): Step, specific to category boundaries within items; Item, constant across steps but specific to items; Common, constant across steps and items; Scale, thresholds with item-specific difficulty and common rating-scale usage across items. For item j and category boundary threshold k, α, common discrimination; α_j, item discrimination; β_{jk}, item-step threshold; δ_j, item difficulty; τ_k, common threshold deviations from the item difficulty.
Note 2 (Models): GRM, graded response model (Samejima, 1969); GRM-C, GRM with a common discrimination parameter; MGRM, modified GRM (Muraki, 1990); MGRM-C, MGRM with a common discrimination parameter; PCM, partial credit model (Masters, 1982); GPCM, generalized PCM (Muraki, 1992); RSM, rating scale model (Andrich, 1978); GRSM, generalized RSM (Muraki, 1992).
Note 3 (Constraints): For models A through D, the variance of the latent trait is fixed at one ($\sigma_\theta^2 = 1$). For models C and D, the sum of the threshold offset parameters is fixed at zero $\sum_{k=1}^{K}\tau_k = 0$.

Table 2.6 Categorical Item Factor Analysis (CIFA) and traditional IRT model specifications.

Model	CIFA∗	IRT	IRT/CIFA parameter correspondence	Constraints
A*/A	$\alpha_j^*\theta + \beta_{jk}^*$	$\alpha_j(\theta - \beta_{jk})$	$\alpha_j = \alpha_j^*$, $\beta_{jk} = -\beta_{jk}^*/\alpha_j$	
B*/B	$\alpha^*\theta + \beta_{jk}^*$	$\alpha(\theta - \beta_{jk})$	$\alpha = \alpha^*$, $\beta_{jk} = -\beta_{jk}^*/\alpha$	
C*/C	$\alpha_j^*\theta + \beta_{jk}^*$	$\alpha_j(\theta - \delta_j - \tau_k)$	$\alpha_j = \alpha_j^*$, $\delta_j = -\beta_{jk}^*/\alpha_j + \tau_k$	$\sum_{k=1}^{K}\tau_k = 0$
D*/D	$\alpha^*\theta + \beta_{jk}^*$	$\alpha(\theta - \delta_j - \tau_k)$	$\alpha = \alpha^*$, $\delta_j = -\beta_{jk}^*/\alpha + \tau_k$	$\sum_{k=1}^{K}\tau_k = 0$

Note: In these models, the latent trait variance is fixed at one ($\sigma_\theta^2 = 1$).

7.4 Model Equivalence, Nesting, and Comparisons

Other commonly described polytomous IRT models are equivalent in fit and prediction to models A through D (see Table 2.7 and Figure 2.9). Among the most widely used are the "Rasch" versions (models B′, B‴, and D′) of models with a common discrimination parameter. In these models, the common discrimination parameter is fixed to 1.0 and the latent trait variance is freely estimated as a model parameter, rather than having its variance fixed to 1.0, as is usual for non-"Rasch" models with a common discrimination parameter. Other useful equivalent models include those (A″ and B″) that decompose the usual threshold parameters β_{jk} into an item location parameter δ_j and associated item-specific threshold offset parameters τ_{jk} which are constrained to sum to zero $\sum_{k=1}^{K}\tau_{jk} = 0$ for each item. In both cases, different model parameters are estimated, but the models remain equivalent in fit and prediction to the base IRT models. Note that models in the last two rows of Table 2.7 necessarily impose so-called rating-scale constraints on the threshold structure of the model. This is relatively explicit for models C and D, but the same constraints apply equally to models C″ and D″, although this is not apparent for their specification given in Table 2.7.

Table 2.7 Some common and useful equivalent models.

Base model	Traditional IRT specification	Equiv. model	"Rasch" version	Equiv. model	Alternate specification	Equiv. model	"Rasch" version location/offset
A	$\alpha_j(\theta - \beta_{jk})$			A″	$\alpha_j(\theta - \delta_j - \tau_{jk})$		
B	$\alpha(\theta - \beta_{jk})$	B′	$\theta - \beta'_{jk}$	B″	$\alpha(\theta - \delta_j - \tau_{jk})$	B‴	$\theta - \delta_j - \tau_{jk}$
C	$\alpha_j(\theta - \delta_j - \tau_k)$			C″	$\alpha_j(\theta - \beta_{jk})$		
D	$\alpha(\theta - \delta_j - \tau_k)$	D′	$\theta - \delta'_j - \tau'_k$	D″	$\alpha(\theta - \beta_{jk})$		

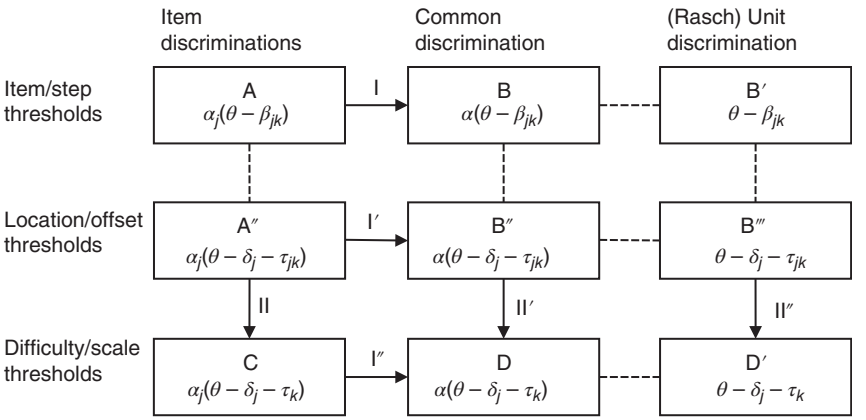


Figure 2.9 Nesting and equivalence relations among common polytomous IRT models.

These models can be arranged in a schematic framework illustrating equivalence, nesting, and some of the basic comparison strategies for the models. The models in Figure 2.9 are arranged in terms of how the discrimination parameters are treated (the columns) and how the threshold parameters are treated (the rows). The discrimination n parameters can be specific to items (α_j, column 1), common across both boundaries and items (α, column 2), or fixed to 1.0 for all items (column 3). Again, the models in the last column of Figure 2.9 are sometimes also referred to as "Rasch" models, but this designation is in practice usually reserved for the models that also use adjacent category comparisons. The threshold parameters can be freely estimated for category boundaries and items (β_{jk}, row 1), transformed into item-specific location and threshold offsets ($\delta_j + \tau_{jk}$, row 2), or have the corresponding threshold offsets equated across items ($\delta_j + \tau_k$, row 3). Because of the equivalence of the models in the first two rows, and the equivalence of the models in the last two columns, there are just four unique models in terms of the number of unique parameters and model fit.

Nested Models can be compared with likelihood ratio model comparisons to test constraints of interest on the model parameters. Within the two main category comparison methods (cumulative probability and adjacent category), a number of informative model comparisons can be conducted to test hypotheses about the model parameters (see Figure 2.9 and Table 2.8). *Model comparisons I, I′, and I″* compare models in the second and third columns of Figure 2.9, and test the improvement in model fit from allowing item-specific discrimination parameters. *Model comparisons II, II′ and II″* compare the last two rows of Table 2.4, and test model fit improvement from allowing item level

Table 2.8 Common one-step model comparisons for the models in Figure 2.9.

Model comparison	Less restricted model		More restricted model		Test of	Conditional on
I	A	$\alpha_j(\theta - \beta_{jk})$	B	$\alpha(\theta - \beta_{jk})$	$\alpha_j = \alpha$	β_{jk}
I′	A″	$\alpha_j(\theta - \delta_j - \tau_{jk})$	B″	$\alpha(\theta - \delta_j - \tau_{jk})$	$\alpha_j = \alpha$	δ_j, τ_{jk}
I″	C	$\alpha_j(\theta - \delta_j - \tau_k)$	D	$\alpha(\theta - \delta_j - \tau_k)$	$\alpha_j = \alpha$	δ_j, τ_k
II	A″	$\alpha_j(\theta - \delta_j - \tau_{jk})$	C	$\alpha_j(\theta - \delta_j - \tau_k)$	$\tau_{jk} = \tau_k$	α_j, δ_j
II′	B″	$\alpha(\theta - \delta_j - \tau_{jk})$	D	$\alpha(\theta - \delta_j - \tau_k)$	$\tau_{jk} = \tau_k$	α, δ_j
II″	B‴	$\theta - \delta_j - \tau_{jk}$	D′	$\theta - \delta_j - \tau_k$	$\tau_{jk} = \tau_k$	δ_j

variation in the threshold offset parameters. Note that the difficulty/scale threshold models in the last row of Figure 2.9 require that the items have the same response categories.

Figure 2.9 and Table 2.8 describe the one-step nesting of adjoining models differing in a single set of parameters, but two-step nesting relations can also be examined jointly. For example, Model D is nested within Model C, which is, in turn, nested within Model A″, leading to a transitive nesting of Model D within Model A″ giving a joint test of the equality of the item discriminations ($\alpha_j = \alpha$) and scale usage ($\tau_{jk} = \tau_k$). Other such nesting relationships can be obtained through following the arrows in Figure 2.9. Nesting along the arrows depicts both parameter and likelihood nesting. With parameter nesting, the more restricted model is obtained by imposing equality or other constraints on the parameters of the less restricted model. Likelihood nesting is based on the fit and prediction of the models, and permits multistep model comparisons of models that also incorporate model equivalence. For example, Model C can be said to be likelihood-nested within Model A because Model C is nested within Model A″, which is in turn likelihood-equivalent to Model A.

8 Four Ordinal Polytomous IRT Models

8.1 Model A: Item Discrimination Parameters and Step Thresholds

Model A is often described mathematically as $\alpha_j(\theta - \beta_{jk})$ with item-specific discrimination parameters α_j, and threshold parameters that are specific to each step of each item β_{jk}. For cumulative probability category comparisons, Model A is the graded response model (GRM), and with adjacent category comparisons, it is the generalized PCM (GPCM). Model A is sometimes alternately represented as the equivalent Model A″ $\alpha_j(\theta - \delta_j - \tau_{jk})$ which estimates item-specific discrimination parameters α_j, and partitions the item-step thresholds $\beta_{jk} = \delta_j + \tau_{jk}$ into item-specific location parameters δ_j and a set of threshold offset parameters τ_{jk} that are constrained to sum to zero $\sum_{k=1}^{K} \tau_{jk} = 0$ for each item.

8.2 Model B: Common Discrimination Parameter and Step Thresholds

Model B can be described mathematically as $\alpha(\theta - \beta_{jk})$, with a discrimination parameter α that is common to all items, and item- and step-specific thresholds β_{jk}. Compared to Model A, Model B constrains the item discrimination parameters $\alpha_j = \alpha$ to be equal. For cumulative probability category comparisons, Model B is the GRM with a common discrimination parameter (GRM-C), and for adjacent category comparisons, it is the PCM. A common equivalent to Model B is the "Rasch"

Model B′ $\theta - \beta_{jk}$, in which the common discrimination parameter is fixed to 1.0, and the variance of the latent trait is freely estimated rather than being fixed to 1.0 as it is in Model B. Model B is sometimes alternatively formulated as the equivalent Model B″ $\alpha(\theta - \delta_j - \tau_{jk})$, including a common discrimination parameter α, an item-specific location parameter δ_j, and a set of threshold offset parameters τ_{jk} that are constrained to sum to zero $\sum_{k=1}^{K} \tau_{jk} = 0$ for each item. Finally, Model B is sometimes also formulated as the equivalent Model B‴ $\theta - \delta_j - \tau_{jk}$ with a unit discrimination and a location-offset formulation of the thresholds.

8.3 Model C: Item Discrimination Parameters with Item Location and Common Threshold Offsets

For items with the same response categories, Model C can be mathematically described as $\alpha_j(\theta - \delta_j - \tau_k)$ with a discrimination parameter α_j and a location or difficulty parameter δ_j for each item, with common category boundary deviations τ_k from the item difficulty estimate where the τ_k are constrained to be equal across items $\tau_{jk} = \tau_k$, and to sum to zero, $\sum_{k=1}^{K} \tau_k = 0$. Model C is the modified GRM (MGRM) for cumulative probability category comparisons and the Generalized Rating Scale Model (GRSM) for adjacent category comparisons.

8.4 Model D: Common Discrimination Parameter with Item Location and Common Threshold Offsets

For items with the same response categories, Model D can be mathematically described as $\alpha(\theta - \delta_j - \tau_k)$ with a common discrimination parameter α, item-specific location or difficulty parameters δ_j, and common category boundary deviations τ_k. Compared to Model C, Model D constrains the item discrimination parameters to be equal $\alpha_j = \alpha$, and compared to Model B″ it constrains the item threshold offsets to be equal $\tau_{jk} = \tau_k$. Model D for cumulative probability comparisons is the MGRM with a common discrimination parameter (MGRM-C), and for adjacent category comparisons, it is the Rating Scale Model (RSM). Model D is sometimes also formulated as Model D' $\theta - \delta_j' - \tau_k'$ with a unit discrimination parameter, item-specific locations and common threshold offsets, with the latent trait variance being freely estimated.

9 Software and Example Analyses

These models, including their equivalent forms, provide a fairly comprehensive, though certainly not exhaustive, description of polytomous IRT models encountered in measurement practice and research. Different software programs for estimating these models can vary in the way they parameterize the models and constrain the model parameters (Childs & Chen, 1999).

As with binary IRT models, we will use Mplus and Stata irt and *gsem* to illustrate these models for polytomous items. The Stata irt procedure is relatively new but provides convenient access to IRT calibration functions for researchers who use Stata for their general-purpose statistical and data analysis work. Stata's *gsem* procedure implements related CIFA functions to estimate a variety of similar models. Mplus is a general statistical and latent variable modeling software program with features that can be useful in IRT model estimation. These models can also be estimated with more specialized IRT software such as Parscale, IRTPro, and flexMIRT, and also with various R packages for IRT analysis. We will use a common data set (Zheng & Rabe-Hesketh, 2007). This dataset, charity.dta, contains five survey questions, ta1 through ta5, measuring faith and trust in charity

organizations. Responses are strongly agree (0), agree (1), disagree (2), and strongly disagree (3). Higher scores indicate higher levels of distrust. The dataset can be obtained with the Stata command *webuse charity*. We use this dataset to present the specification of the models, organized by the type of category comparisons (cumulative probability vs. adjacent category), the software program used, and the model estimated.

10 Cumulative Probability Models Using Stata

10.1 Model A: The Graded Response Model (GRM)

With cumulative probability category comparisons, Model A (the GRM) can be obtained by using the following Stata *irt* syntax. Specification of the GRM in Stata is accomplished with the *grm* specification of the *irt* procedure, yielding the standard IRT Model A parameters $\alpha_j(\theta - \beta_{jk})$. The *estat ic* command is used to obtain the model log-likelihood, number of model parameters, and information criteria (AIC and BIC) for comparatively judging model fit. As before, comments preceded by two slashes (//) are optional and are used here to annotate the input. The Stata *irt grm* command can be followed by the *irtgraph* command to plot item category response functions and item and test information functions, and also by the *estat report* command to a custom format, label, and display parameter estimates. Stata *irt* gives estimates, standard errors, and confidence intervals for all model parameters, and also provides significance tests for discrimination parameters.

```
*** Cumulative probability Model A: Graded response model (GRM) Stata irt
irt grm ta1-ta5  // GRM

*** Example post-estimation statistics and plots
estat ic             // Information Statistics (LL, AIC, BIC)
irtgraph icc ta1     // Item characteristic curve (ICC) for item ta1
estat report, byparm // Condensed report of the IRT parameters
```

The parameters of the GRM can also be estimated using *gsem* in Stata. In this case, the factor F designates the latent trait, the *ologit* option specifies the appropriate response function for cumulative probability comparisons, and the *variance* option constrains the variance of the latent trait to 1. The *forvalue* structure loops over the five items and then over the three steps to obtain the GRM IRT threshold parameters for each item/step combination using the *nlcom* command.

```
*** Cumulative probability Model A: Graded response model (GRM) Stata gsem
gsem (F->ta1-ta5), ologit variance(F@1)        // Estimate the CIFA model

*** Transform from CIFA to IRT parameterization
forvalue j=1/5 {                                 // Items
 forvalue k=1/3 {                                // Steps
  nlcom (b`j'`k': _b[/ta`j':cut`k']/_b[ta`j':F]) // IRT threshold
parameters
 }
}
```

The *gsem* estimation of the CIFA GRM provides the same model fit information and will produce the same latent trait person ability estimates (not requested here) as the *irt* estimation of the GRM. However, the *irt* command can be followed by the *irtgraph* and *estat report* commands to better visualize and display the model results.

10.2 Model B: The Graded Response Model with Common Discrimination (GRM-C)

Special parameter constraints can be used to equate the discrimination parameters using with Stata *irt* to estimate Model B (GRM-C), yielding the standard Model B parameters $\alpha(\theta - \beta_{jk})$. The *cns (a@a)* option here equates the discrimination parameter across items.

```
*** Cumulative probability model B: Graded response model
*** with a common discrimination parameter (GRM-C) Stata irt
irt grm ta1-ta5, cns(a@a) // GRM-C
```

The parameters of the GRM-C can also be obtained via *gsem* in Stata. Compared to Model A, the designation *@a* following the list of indicators constrains the factor loading or discrimination parameter to be the same for all items.

```
*** Cumulative probability Model B: [GRM-C] Stata gsem
gsem (F->ta1-ta5@a), ologit variance(F@1)          // Estimate CIFA model

*** Transform from CIFA to IRT parametrization
forvalue j=1/5 {                                    // Items
 forvalue k=1/3 {                                   // Steps
 nlcom (b`j'`k': _b[ta`j'_cut`k':_cons]/_b[ta1:F]) // IRT threshold
parameters
 }
}
```

Neither *irt* nor *gsem* in Stata can be used to directly estimate the parameters of the traditional IRT Models C or D for cumulative probability category comparisons because the standard irt constraint functions are not able to impose the required constraints on the model parameters.

11 Adjacent Category Models Using Stata

11.1 Model A: The Generalized Partial Credit Model (GPCM)

The GPCM can be obtained by specifying *gpcm* on the Stata *irt* command as shown here, yielding the standard Model A parameters $\alpha_j(\theta - \beta_{jk})$.

```
***Adjacent category Model A: Generalized partial credit model (GCPM)
irt gpcm ta1-ta5 // GCPM
```

11.2 Model B: The Partial Credit Model (PCM)

The PCM can similarly be obtained by specifying pcm on the Stata irt command, yielding the standard Model B parameters $\alpha(\theta - \beta_{jk})$.

```
***Adjacent category Model B: Partial credit model (PCM) Stata irt
irt pcm ta1-ta5 // PCM
```

The PCM can also be obtained as a GPCM with a common discrimination parameter.

```
irt gpcm ta1-ta5, cns(a@a) // PCM as a GPCM with a common
discrimination
```

11.3 Model C: The Generalized Rating Scale Model (GRSM)

Stata *irt* does not provide a way to directly estimate Model C, but does have an available specification for Model D, the RSM.

11.4 Model D: The Rating Scale Model (RSM)

The RSM can be obtained by specifying *rsm* on the Stata *irt* command. The RSM estimated by Stata irt is actually a constrained version of the PCM that estimates category boundary thresholds like those of Model B, $\alpha(\theta - \beta_{jk})$, treating the first threshold of each item as a "location parameter," and then separately equating the differences between the first threshold and each of the remaining thresholds across all of the items.

```
***Adjacent Category Model D: Rating scale model [RSM] Stata irt
irt rsm ta1-ta5 // RSM
```

The *gsem* procedure in Stata is not able to easily estimate the Models A through D for adjacent category comparisons because *gsem* does not have an adjacent category ordinal response function that operates for PCMs in the way that ologit does for GRMs. However, in principle, it should be possible to reparameterize these as special cases of the Nominal Response Model (NRM). Indeed, all the adjacent category models estimated by Stata irt internally use specially constrained *gsem* multinomial NRM models.

11.5 Summary

For cumulative probability category comparisons, Stata *irt* and *gsem* can be used to estimate the usual IRT parameters for Model A (the GRM) and Model B (the GRM-C), but not model C (MGRM) or model D (MGRM-C). For adjacent category comparisons, Stata *irt* can be used to estimate Models A (the GPCM), B (the PCM) and D (the RSM). Stata does not provide a way to directly estimate Model C for either cumulative probability comparisons (the MGRM) or adjacent category comparisons (the GRSM).

12 Cumulative Probability Models Using Mplus

12.1 Model A: The Graded Response Model (GRM)

With cumulative probability category comparisons, Model A (the GRM) can be obtained by using the following Mplus syntax. If one only wanted the CIFA slope and intercept parameters $\alpha_j^*\theta + \beta_{jk}^*$ along with model fit and prediction, the first seven lines (through the MODEL command) would be sufficient. However, it is also possible to obtain the traditional IRT discrimination and threshold parameters $\alpha_j(\theta - \beta_{jk})$ by labeling the original CIFA intercept parameters, and using the accompanying MODEL CONSTRAINT: specification to transform the CIFA intercepts into the IRT thresholds.

Because Mplus, like Stata *gsem*, is a CIFA based program, it begins by defining a latent trait or factor we have called F, whose variance is fixed to 1 (with its mean being fixed to 0 by default). The MODEL: command specifies that the factor loadings or slopes for all of the items are to be freely estimated by using the asterisk (*) following the item list, and the corresponding CIFA slopes α_j^* are given labels *aj* that are used later to calculate the IRT threshold parameters. The CIFA intercepts β_{jk}^* are also labeled with the convention c*k_j* indicating the intercept for category boundary *k* and item *j*.

```
Title: Cumulative probability Model A: Graded response model [GRM]
Data: File is charity.dat;   ! Data file
Variable: Names are ta1-ta5; ! Variable names
  Missing are all (-9999);   ! Missing values
  CATEGORICAL ARE ALL;       ! Variables are categorical
Analysis: ESTIMATOR IS ML;   ! Maximum likelihood estimation
MODEL: f@1; f BY ta1-ta5* (a1-a5);   ! Model A GRM
 [ta1$1-ta5$1] (c1_1-c1_5);          ! Label intercepts
 [ta1$2-ta5$2] (c2_1-c2_5);          !
 [ta1$3-ta5$3] (c3_1-c3_5);          !
MODEL CONSTRAINT:                    ! Model constraints
 NEW(b1_1-b1_5 b2_1-b2_5 b3_1-b3_5); ! Get IRT thresholds
 DO ($,1,5) DO (#,1,3) b#_$=c#_$/a$; !
```

The IRT thresholds β_{jk} are designated *bk_j* in the code and are defined as new parameters on the MODEL CONSTRAINT: command. Mplus' DO commands are used to obtain the thresholds by dividing the intercepts by the corresponding discrimination parameters. For convenience in specifying the syntax using variable lists in Mplus, notice that the order of the item *j* and category boundary *k* subscripts for intercepts and thresholds have been reversed and are separated by an underscore character "_". In these DO commands, items are indexed with the dollar sign ($) and category boundaries are indexed with the pound sign (#). For the GRM, the IRT discrimination parameters are equal to the corresponding Mplus CIFA factor loadings $\alpha_j = \alpha_j^*$, and because the Mplus intercepts are the negative of the usual CIFA intercepts, the IRT threshold parameters are equal to the intercepts divided by the discrimination parameters $\beta_{jk} = \beta_{jk}^*/\alpha_j$. Because these are derived parameters, they do not affect the estimation or fit of the original model.

It is also possible to derive the alternately specified IRT location and threshold offset parameters $\delta_j - \tau_{jk}$ from the IRT threshold parameters β_{jk} by appending the following code snippet to the

previous Mplus Model A syntax. This augments the MODEL CONSTRAINT: command with new parameters that are analogous to the location (d1–d5) and threshold offset (e1_1-e1_5 e2_1-e2_5 e3_1-e3_5) parameters found in other models such as Model C and Model D. The location parameters are conveniently derived as the average of the IRT thresholds for an item, and the offsets are obtained as the deviation of the thresholds from the item location parameter. Again, since these are derived parameters, they do not affect the estimation or fit of the original model.

```
...
NEW(d1-d5 e1_1-e1_5 e2_1-e2_5 e3_1-e3_5);  ! Locations and offsets
 DO ($,1,5) d$=(b1_$+b2_$+b3_$)/3;         ! Item locations
 DO ($,1,5) DO (#,1,3) e#_$=b#_$-d$;       ! Threshold offsets
```

12.2 Model B: The Graded Response Model with Common Discrimination

With cumulative category comparisons, Model B is the GRM-C which can be estimated using the following MPLUS syntax, yielding the CIFA slope and intercept parameters $\alpha^*\theta + \beta^*_{jk}$, and the IRT discrimination and threshold parameters $\alpha(\theta - \beta_{jk})$ by labeling the CIFA intercept parameters and using appropriate MODEL CONSTRAINT: commands. The GRM-C differs from the GRM in that the slopes are equated by giving them a common label (a) in the MODEL: section, which is subsequently also used with the NEW and DO commands in the MODEL CONSTRAINT: section.

```
TITLE: Mplus CIFA Cumulative Probability Model B: [GRM-C]
DATA: FILE IS charity_flexmirt.csv; ! Data file
VARIABLE: NAMES ARE ta1-ta5;        ! Variable names
 MISSING ARE ALL (-9);              ! Missing values
 CATEGORICAL ARE ALL;               ! Cumulative probability ordinal data
ANALYSIS: ESTIMATOR IS ML;          ! Maximum likelihood estimation
MODEL: F@1; F BY ta1-ta5* (a);      ! Model

[ta1$1-ta5$1] (c1_1-c1_5);          ! Label the intercept parameters
[ta1$2-ta5$2] (c2_1-c2_5);          !
[ta1$3-ta5$3] (c3_1-c3_5);          !
MODEL CONSTRAINT:                   ! Model Constraints
NEW(b1_1-b1_5 b2_1-b2_5 b3_1-b3_5); ! Get reproduced IRT thresholds
DO ($,1,5) DO (#,1,3) b#_$=c#_$/a;  !
```

For the GRM-C, the common IRT discrimination parameter is equal to the common Mplus CIFA factor loading $\alpha = \alpha^*$, and the IRT threshold parameters are equal to the CIFA intercepts divided by the common discrimination parameter, $\beta_{jk} = \beta^*_{jk}/\alpha$.

12.3 Model C: The Modified Graded Response Model (MGRM)

For cumulative probability comparisons, Model C is the MGRM, which can be estimated with the following Mplus syntax, yielding the (implicitly constrained) CIFA slope and intercept parameters $\alpha^*_j\theta + \beta^*_{jk}$, along with the appropriate IRT item discrimination, item location, and common

threshold offset parameters $\alpha_j(\theta - \delta_j - \tau_k)$ where the offsets are constrained to average zero $\sum_{k=1}^{K} \tau_k = 0$. This syntax differs from the GRM syntax by placing constraints on the IRT thresholds using the MODEL CONSTRAINT: to obtain the IRT specification of the model coinciding with the implementation in the program Parscale which places constraints directly on the IRT threshold offsets rather than on corresponding CIFA intercept offsets. It was necessary to supply some "judicious" starting values for the location and threshold offset parameters in order to get the model to converge successfully and impose the appropriate constraints on the model parameters. A second segment of the MODEL CONSTRAINT: command was used to derive the standard IRT threshold parameters as well as the locations and threshold offsets.

```
TTitle: Cumulative probability Model C: Modified graded response model (MGRM)
Data: File is charity.dat ;
Variable: Names are ta1-ta5;
  Missing are all (-9999);
  CATEGORICAL ARE ALL;
Analysis: ESTIMATOR IS ML;
MODEL: f@1; f BY ta1-ta5* (a1-a5);       ! Item-specific discriminations
 [ta1$1-ta5$1] (c1_1-c1_5);              ! Label the intercept parameters
 [ta1$2-ta5$2] (c2_1-c2_5);              !
 [ta1$3-ta5$3] (c3_1-c3_5);              !
MODEL CONSTRAINT:                        ! IRT location and offset
parameters
 NEW(d1-d5*1 t1*-1 t2*0 t3*1);           ! "Judicious" starting values
 DO ($,1,5) DO (#,1,3) c#_$=a$*d$+a$*t#; ! Model constraints
 t1=0-t2-t3;                             ! Constrain offsets to average 0
 NEW(b1_1-b1_5 b2_1-b2_5 b3_1-b3_5);     ! Obtain IRT threshold parameters
 DO ($,1,5) DO (#,1,3) b#_$=c#_$/a$;     !
```

The CIFA intercept structure of the items is re-parameterized in the MODEL CONSTRAINT: section to yield a location parameter δ_j for each item and a threshold offset τ_k for each category boundary, $\delta_j = -\beta_{jk}^*/\alpha_j + \tau_k$, with the offsets constrained to average zero $\sum_{k=1}^{K} \tau_k = 0$. Notice that this differs slightly from the corresponding specification in Table 2.3, due to jointly estimating the IRT item location and common category boundary offset parameters with Mplus.

12.4 Model D: The Modified Graded Response Model with Common Threshold Offsets (MGRM-C)

With cumulative probability comparisons, Model D is the MGRM-C, which can be estimated with the following Mplus syntax, yielding the CIFA slope and intercept parameters $\alpha_j^*\theta + \beta_{jk}^*$ and the traditional IRT common discrimination, item location, and common threshold offset parameters $\alpha(\theta - \delta_j - \tau_k)$, corresponding to the model estimated by Parscale. This model differs from the Model C MGRM in that it estimates and uses a common discrimination parameter (a). In addition to the location and threshold offset parameters, the syntax below also obtains the standard IRT threshold discrimination parameters subject to the model constraints.

```
Title: Cumulative probability Model D: Modified graded response
 model with a common discrimination parameter [MGRM-C]
Data: File is charity.dat ;
Variable: Names are ta1-ta5;
  Missing are all (-9999);
  CATEGORICAL ARE ALL (GPCM);
Analysis: ESTIMATOR IS ML;
MODEL: f@1; f BY ta1-ta5* (a);          ! Model D MGRM-C
 [ta1$1-ta5$1] (c1_1-c1_5);             ! Label the intercept parameters
 [ta1$2-ta5$2] (c2_1-c2_5);             !
 [ta1$3-ta5$3] (c3_1-c3_5);             !
MODEL CONSTRAINT: NEW(d1-d5 t1-t3);     ! IRT location and offset
parameters
DO ($,1,5) DO (#,1,3) c#_$=a*(d$+t#);   ! Model constraints (unscaled)
t1=0-t2-t3;                             ! Constrain offsets to average 1
NEW(b1_1-b1_5 b2_1-b2_5 b3_1-b3_5);     ! Obtain IRT threshold parameters
DO ($,1,5) DO (#,1,3) b#_$=c#_$/a;      !
```

The MGRM-C has a common discrimination parameter α, along with item location parameters δ_j and category boundary threshold offsets τ_k, related with the expression $\delta_j = -\beta^*_{jk}/\alpha + \tau_k$, where the offsets constrained to average zero $\sum_{k=1}^{K} \tau_k = 0$.

Models A, B, C, and D for cumulative probability comparisons can be seen as variants of the unidimensional GRM and can be obtained with Mplus by (i) declaring the ordinal polytomous items to be CATEGORICAL variables, (ii) estimating a single-factor CIFA model with the factor variance fixed to one, (iii) using appropriate parameter labels to the CIFA intercepts, and (iv) applying the needed constraints and transformations to obtain appropriate IRT threshold or location and offset parameters.

13 Adjacent Category Models with Mplus

13.1 Model A: The Generalized Partial Credit Model (GPCM)

With adjacent category comparisons, the GPCM item location and threshold offset parameters (Model A″) can be obtained by using the following Mplus syntax.

```
Title: Adjacent category Model A: Generalized partial credit model [GPCM]
Data: File is charity.dat ;
Variable: Names are ta1-ta5;
  Missing are all (-9999);
  CATEGORICAL ARE ALL (GPCM); ! Adjacent category ordinal data
Analysis: ESTIMATOR IS ML;    ! Maximum likelihood estimation
MODEL: f@1; f BY ta1-ta5*;    ! Item-specific discriminations
```

The critical element for adjacent category comparisons is the use of the (GPCM) specification to request adjacent category comparisons on the CATEGORICAL ARE ALL statement. In addition to the discrimination parameters, Mplus automatically produces both the CIFA intercepts and IRT item location and threshold offset parameters for adjacent category models. By also labeling the CIFA intercepts and applying the appropriate transformations, Mplus can produce the traditional Model A threshold parameters $\alpha_j(\theta - \beta_{jk})$ in addition to Mplus' default location and offset parameters for Model A″ $\alpha_j(\theta - \delta_j - \tau_{jk})$.

```
...
 [ta1$1-ta5$1] (c1_1-c1_5);           ! Label the intercept parameters
 [ta1$2-ta5$2] (c2_1-c2_5);           !
 [ta1$3-ta5$3] (c3_1-c3_5);           !
MODEL CONSTRAINT:
NEW(b1_1-b1_5 b2_1-b2_5 b3_1-b3_5); ! Obtain IRT boundary
thresholds
DO ($,1,5) DO (#,1,3) b#_$=c#_$/a$; !
```

Although labeling the intercepts and transforming from CIFA intercepts to IRT thresholds is optional, it can be helpful when comparing the Mplus results to those of programs like Stata which gives the derived traditional IRT thresholds but does not provide the associated location and threshold offset parameters.

13.2 Model B: The Partial Credit Model (PCM)

The PCM (Model B″) $\alpha(\theta - \delta_j - \tau_{jk})$ can be obtained by requesting a common slope/discrimination parameter (a) on the MODEL: command.

```
Title: Adjacent category Model B: Partial credit model [PCM]
Data: File is charity.dat ;
Variable: Names are ta1-ta5;
 Missing are all (-9999);
 CATEGORICAL ARE ALL (GPCM);   ! Adjacent category logits
Analysis: ESTIMATOR IS ML;     ! Maximum likelihood estimation
MODEL: f@1; f BY ta1-ta5* (a); ! Common Discrimination
```

As with Model A (the GPCM), it is possible to obtain the derived IRT threshold parameters by labeling the CIFA intercepts and using the MODEL CONSTRAINT: command to transform the CIFA intercepts into the corresponding IRT thresholds. This differs from Model A in that there is a common discrimination parameter rather than item-specific discriminations.

```
...
 [ta1$1-ta5$1] (c1_1-c1_5);          ! Label the intercept parameters
 [ta1$2-ta5$2] (c2_1-c2_5);          !
 [ta1$3-ta5$3] (c3_1-c3_5);          !
NEW(b1_1-b1_5 b2_1-b2_5 b3_1-b3_5); ! Obtain IRT threshold parameters
DO ($,1,5) DO (#,1,3) b#_$=c#_$/a; !
```

13.3 Model C: The Generalized Rating Scale Model (GRSM)

For adjacent category comparisons, the GRSM (Model C) can be obtained by equating the threshold differences of the GPCM to equal across items, as shown in the following syntax example. This involves labeling the CIFA intercept parameters and imposing rating scale constraints on the thresholds using the MODEL CONSTRAINT: command.

```
Title: Adjacent category Model C: Generalized Rating Scale Model [GRSM]
Data: File is charity.dat ;
Variable: Names are ta1-ta5;
  Missing are all (-9999);
  CATEGORICAL ARE ALL (GPCM);
Analysis: ESTIMATOR IS ML;
MODEL: f@1; f BY ta1-ta5*1 (a1-a5);      ! Item-specific discriminations
 [ta1$1-ta5$1] (c1_1-c1_5);              ! Label the intercept parameters
 [ta1$2-ta5$2] (c2_1-c2_5);              !
 [ta1$3-ta5$3] (c3_1-c3_5);              !
MODEL CONSTRAINT: NEW(d1-d5 t1-t3);      ! IRT location and offset
                                           parameters
 DO ($,1,5) DO (#,1,3) c#_$=a$*d$+a$*t#; ! Model constraints
 t1=0-t2-t3;                             ! Constrain offsets to average 0
 NEW(b1_1-b1_5 b2_1-b2_5 b3_1-b3_5);     ! Obtain IRT threshold parameters
 DO ($,1,5) DO (#,1,3) b#_$=c#_$/a$;     !
```

For completeness and possible comparisons, the estimated (and constrained) CIFA intercepts are transformed into corresponding IRT threshold parameters.

13.4 Model D

With adjacent category comparisons, the traditional RSM also uses constraints on the MODEL: command to estimate a common slope/discrimination parameter. Again, Mplus automatically provides the appropriately constrained Model D item location and threshold offset parameters.

```
T Title: Adjacent category Model D: Rating Scale Model [RSM]
Data: File is charity.dat ;
Variable: Names are ta1-ta5;
  Missing are all (-9999);
  CATEGORICAL ARE ALL (GPCM);
Analysis: ESTIMATOR IS ML;
MODEL: f@1; f BY ta1-ta5*1 (a);          ! Common discrimination parameter
 [ta1$1-ta5$1] (c1_1-c1_5);              ! Label the intercept parameters
 [ta1$2-ta5$2] (c2_1-c2_5);              !
 [ta1$3-ta5$3] (c3_1-c3_5);              !
MODEL CONSTRAINT: NEW(d1-d5 t1-t3);      ! IRT location and offset parameters
 DO ($,1,5) DO (#,1,3) c#_$=a*(d$+t#); ! Model constraints (unscaled)
 t1=0-t2-t3;                             ! Constrain offsets to average 1
 NEW(b1_1-b1_5 b2_1-b2_5 b3_1-b3_5);     ! Obtain IRT threshold parameters
 DO ($,1,5) DO (#,1,3) b#_$=c#_$/a;      !Equate across items
```

Again, Mplus' MODEL CONSTRAINT: command is used to transform the CIFA intercepts from the model with common discrimination and common threshold offsets to corresponding standard IRT thresholds.

14 Model Implementation

Within each category comparison method, the four main models in Table 2.9 (A through D) can be estimated and compared using Stata (irt and *gsem*) and Mplus.

14.1 Cumulative Probability Models

Stata *irt* can directly estimate the GRM and GRM-C (models A and B), giving the standard IRT threshold parameters along with the appropriate discrimination parameters. Stata *gsem* can also estimate these models based on the CIFA slope and intercept parameterization (models A* and B*) followed by transformations with Stata's *nlcom* command to produce the IRT threshold parameters for the two models. Mplus can estimate the GRM, GRM-C, MGRM, and MGRM-C based on the CIFA parameterization (models A*, B*, C*, and D*). With appropriate model constraints and parameter transformations, the corresponding IRT threshold parameters and equivalent IRT location and threshold offset parameters can be produced as shown in this guide.

14.2 Adjacent Category Models

Stata *irt* can directly estimate the GPCM, PCM, and RSM (models A, B, and D), producing the standard IRT threshold parameters. Mplus estimates the GPCM, PCM, GRSM, and RSM based on the CIFA parameterization (models A*, B*, C*, and D*), producing the IRT location and threshold offset parameters appropriate to each model. Subsequent transformations can also be used to produce the corresponding standard IRT threshold parameters applicable to a given

Table 2.9 Ordinal polytomous IRT models estimable with Stata and Mplus.

Model		Stata		
Cumulative probability		**irt**	***gsem***	**Mplus**
A	GRM	A	A*, (A)	A*, (A)
B	GRM-C	B	B*, (B)	B*, (B)
C	MGRM			C*, (C)
D	MGRM-C			D*, (D)
Adjacent category				
A	GPCM	A		A*, A″
B	PCM	B		B*, B″
C	GRSM			C*, C
D	RSM	D		D*, D

Note: GRM = Graded Response Model; GRM-C = Graded Response Model with a Common discrimination parameter; MGRM = Modified Graded Response Model; MGRM-C = Modified Graded Response Model with a Common discrimination parameter; GPCM = Generalized Partial Credit Model; PCM = Partial Credit Model; GRSM = Graded Rating Scale Model; RSM = Rating Scale Model.

model. The GPCM has item-specific discrimination parameters and a saturated threshold structure where there is a location parameter and a set of threshold offset parameters that sum to zero for each item. The PCM has a common discrimination parameter for all the items but, like the GPCM, has a saturated location and threshold offset structure. The GRSM has item-specific discrimination parameters, but imposes so-called "rating scale" constraints on the thresholds such that each item has an overall location or difficulty parameter which is accompanied by corresponding threshold offsets for each category boundary that are equal across all of the items. In addition to these rating scale constraints, the RSM also estimates a common discrimination parameter.

Although many of these models can also be estimated with more specialized IRT software such as Parscale, IRTPro, or flexMIRT, we have demonstrated their availability and utility in a general purpose statistical and data analysis package like Stata and in a general statistical modeling program like Mplus. Psychometric professionals and experts can take advantage of the additional detail and unique capabilities of dedicated IRT software, but many working researchers may be more familiar and comfortable with software like Stata that they use for preparing and managing their research data, or with software like Mplus that they may also use for many other statistical modeling and analysis tasks. Although Stata irt is basically limited to univariate IRT models, *gsem* in Stata, like Mplus, has multidimensional, multilevel, mixture, and other more general statistical modeling possibilities. The introduction provided in this guide can serve as a foundation or starting point for pursuing multidimensional and explanatory IRT models and integrating IRT analyses into other multilevel and latent class type models. For many common models, Stata irt has a very concise and intuitive syntax that is well-suited to researchers who currently use Stata elsewhere in their work. Some users may also find it convenient to use from Stata's graphical menu-driven interface.

Stata, Mplus, and the specialized IRT software mentioned earlier are all commercial products, and therefore have a substantial monetary cost for those who do not already have access to them. One is unlikely to purchase Stata or Mplus just for doing IRT modeling. In passing, we do note that the exercises in this guide, as with Guide I on CTT, can be accomplished with the demo version of Mplus. For real research work, however, one would likely need to acquire or gain access to Stata, Mplus, or one of the specialized IRT software programs. So far, we have scrupulously avoided demonstration or discussion of open-source or other freely available IRT software, or of other related software with IRT-like capabilities. One could envision a version of this guide using such software, but we leave that for another day.

15 Model Fit for the Example Data

Table 2.10 summarizes the results of applying four cumulative probability models (GRM, GRM-C, MGRM, and MGRM-C) and four adjacent category models (GPCM, PCM, GRSM, and RSM) to the Charity data. The table presents the −2 log-likelihood values estimated by Mplus and by Stata irt and *gsem*.

Both Stata and Mplus give similar model fit results for the GRM, GRM-C, GPCM, and PCM. For both cumulative probability and adjacent category comparisons, Model A and Model B are readily estimated by both Stata and Mplus. Mplus estimates all four of the models for both kinds of category comparisons.

Table 2.10 Number of parameters and −2 log-likelihood values for selected polytomous IRT models and software for the Charity data.

Model			Stata		
Cumulative probability		**# Parm.**	**irt**	***gsem***	**Mplus**
A	GRM	20	10318.56	10318.56	10318.56
B	GRM-C	16	10373.70	10373.70	10373.70
C	MGRM	12			10458.55
D	MGRM-C	8			10534.55
Adjacent category					
A	GPCM	20	10353.81		10353.81
B	PCM	16	10419.17		10419.17
C	GRSM	12			10484.52
D	RSM	8	10587.86		10587.86

Note: GPCM, generalized partial credit model; GRM, graded response model; GRM-C, graded response model with a common discrimination parameter; GRSM, graded rating scale model; MGRM, modified graded response model; MGRM-C, modified graded response model with a common discrimination parameter; PCM, partial credit model; RSM, rating scale model.

A. Appendix

Bock & Lieberman Data:

1 1 1 1 1 3
1 1 1 1 2 6
1 1 1 2 1 2
1 1 1 2 2 11
1 1 2 1 1 1
1 1 2 1 2 1
1 1 2 2 1 3
1 1 2 2 2 4
1 2 1 1 1 1
1 2 1 1 2 8
1 2 1 2 1 0
1 2 1 2 2 16
1 2 2 1 1 0
1 2 2 1 2 3
1 2 2 2 1 2
1 2 2 2 2 15
2 1 1 1 1 10
2 1 1 1 2 29
2 1 1 2 1 14
2 1 1 2 2 81
2 1 2 1 1 3
2 1 2 1 2 28
2 1 2 2 1 15

2 1 2 2 2 80
2 2 1 1 1 16
2 2 1 1 2 56
2 2 1 2 1 21
2 2 1 2 2 173
2 2 2 1 1 11
2 2 2 1 2 61
2 2 2 2 1 28
2 2 2 2 2 298

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Andrich, D. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, *34*, 8–14.

Bacci, S., Bartolucci, F., & Gnaldi, M. (2014). A class of Multidimensional Latent Class IRT models for ordinal polytomous item responses. *Communications in Statistics-Theory and Methods*, *43*, 787–800.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179–197

Childs, R. A., & Chen, W. H. (1999). Software note obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement*, *23*, 371–379.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Earlbaum Associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*, 59–71.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*, 36–48.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(S1), 100–114.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.

Zheng, X., & Rabe-Hesketh, S. (2007). Estimating parameters of dichotomous and ordinal item response models with gllamm. *Stata Journal*, *7*, 313–333.

http://www.shinyitemanalysis.org/
https://cemp.shinyapps.io/ShinyItemAnalysis/

Computational Guide 3

Factor Analysis

CHAPTER MENU

1 Unidimensional and Multidimensional Factor Analysis, 438
- 1.1 Factor Analysis with Multiple Factors, 438
- 1.2 Exploratory and Confirmatory Factor Analysis, 439
- 1.3 The Independent Clusters Correlated Factors CFA Model, 440
- 1.4 The Correlated Factors Exploratory Factor Analysis (EFA) Model, 441
- 1.5 Higher-Order Factor Models, 441
- 1.6 Bifactor Models, 442
- 1.7 Data Example, 444

2 Correlated Factors Model, 445
- 2.1 The Correlated Factors Exploratory Factor Analysis (EFA) Model, 445
- 2.2 Communalities and Exploratory Structural Equation Modeling in Mplus, 446
- 2.3 The Correlated Factors Confirmatory Factor Analysis (CFA) Model, 448

3 Higher-Order Factor Models, 448
- 3.1 The Higher-Order Factor CFA, 448
- 3.2 The Higher Order CFA Model with Direct Indicators of the Higher Order Factor, 450
- 3.3 The Higher Order EFA/ESEM Model with Direct Indicators of the Higher Order Factor, 451

4 Bifactor Models, 452
- 4.1 The Symmetric or Complete Bifactor CFA, 452
- 4.2 The Bifactor CFA with a Reference Factor, 453
- 4.3 The Bifactor CFA Model with Direct Indicators of the General Factor, 454
- 4.4 The Orthogonal Bifactor EFA/ESEM with Direct Indicators of the General Factor, 455
- 4.5 Model Fit of Multidimensional Factor Analysis Models, 456

5 The Relationship Between the Higher Order and Bifactor Models, 457
- 5.1 The Higher Order Model as a Constrained Bifactor Model, 457
- 5.2 Loadings of the Indicators on the Higher-Order Factor Using the Higher Order Model, 458
- 5.3 The Bifactor Model as an Extended Higher Order Factor Model, 461

6 Reliability Estimation for Multidimensional Constructs, 462
- 6.1 Stratified Alpha, 462
- 6.2 Reliability Index with the Correlated Factors Model, 464
- 6.3 Reliability Index for the Higher Order Factor Model, 466
- 6.4 Reliability Estimation Via the Reliability Index in the Bifactor CFA, 467
- 6.5 Reliability Estimation for Higher-Order Factor Models Using Analytic Formulas, 469

7 Omega Reliability Estimation Via Reliability Formulas for Hierarchical Data, 471
- 7.1 Omega Hierarchical, 471
- 7.2 Omega Total, 471
- 7.3 Omega Hierarchical Subscale, 472
- 7.4 Omega Subscale, 472
- 7.5 Calculating Omega Reliability Coefficients from Factor Loadings and Residual Variances, 472

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

8 Expected Common Variance, 475
8.1 Expected Common Variance for the General Factor, 475
8.2 Expected Common Variance for Subscales, 475
8.3 Item Expected Common Variance, 475
9 Other Reliability Related Information, 477
9.1 Construct Replicability Index, 477
9.2 Average Relative Parameter Bias, 477
9.3 Factor Determinacy, 478
10 Special Situations in Reliability Assessment, 480
10.1 Negative Factor Loadings on the Specific Factors, 480
10.2 Correlated Indicator Residuals, 481
References, 482
Online Datasets and Services, 483

Classical Test Theory (CTT) and Item Response Theory (IRT) have customarily had a quite strong assumption of unidimensionality. In unidimensional CTT and IRT, a construct is modeled by a single latent variable or latent trait, typically accompanied by the creation of a single associated composite score, or score estimate, for each respondent or examinee. The assumption and application of unidimensionality has been a basic foundation of many developments in measurement. Still, many of the domains and concepts for which measures are sought are irreducibly multidimensional or multifaceted. Sometimes associated phenomena are inextricably bound up together, and their measurement must accommodate these inherent connections. In the following, we use the metaphor and practice of factor analysis to outline basic aspects of CTT and IRT for multidimensional data.

1 Unidimensional and Multidimensional Factor Analysis

1.1 Factor Analysis with Multiple Factors

Sometimes the observed measures can be jointly influenced by multiple latent variables. If there are nine items, subtests, or indicators, with three latent variables or factors, the usual factor analysis model can be expressed in matrix form as:

$$\mathbf{Y} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \\ \nu_5 \\ \nu_6 \\ \nu_7 \\ \nu_8 \\ \nu_9 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \\ \lambda_{41} & \lambda_{42} & \lambda_{43} \\ \lambda_{51} & \lambda_{52} & \lambda_{53} \\ \lambda_{61} & \lambda_{62} & \lambda_{63} \\ \lambda_{71} & \lambda_{72} & \lambda_{73} \\ \lambda_{81} & \lambda_{82} & \lambda_{83} \\ \lambda_{91} & \lambda_{92} & \lambda_{93} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{bmatrix}$$

Here, y_j is a person's observed score on indicator j, ν_j is the intercept for indicator j, λ_{jk} is the factor loading for indicator j on factor η_k, and ε_j is the person's indicator j residual. This general model applies to both exploratory and confirmatory factor analyses. The associated model implied covariance matrix of the observed measures can be expressed in matrix form as

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$$

$$= \begin{bmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \\ \lambda_{41} & \lambda_{42} & \lambda_{43} \\ \lambda_{51} & \lambda_{52} & \lambda_{53} \\ \lambda_{61} & \lambda_{62} & \lambda_{63} \\ \lambda_{71} & \lambda_{72} & \lambda_{73} \\ \lambda_{81} & \lambda_{82} & \lambda_{83} \\ \lambda_{91} & \lambda_{92} & \lambda_{93} \end{bmatrix} \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi_{33} \end{bmatrix} \begin{bmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} & \lambda_{41} & \lambda_{51} & \lambda_{61} & \lambda_{71} & \lambda_{81} & \lambda_{91} \\ \lambda_{12} & \lambda_{22} & \lambda_{32} & \lambda_{42} & \lambda_{52} & \lambda_{62} & \lambda_{72} & \lambda_{82} & \lambda_{92} \\ \lambda_{13} & \lambda_{23} & \lambda_{33} & \lambda_{43} & \lambda_{53} & \lambda_{63} & \lambda_{73} & \lambda_{83} & \lambda_{93} \end{bmatrix} + \boldsymbol{\Theta}$$

$\boldsymbol{\Psi}$ is a $k \times k$ symmetric covariance matrix of the latent variables, and $\boldsymbol{\Theta}$ is a $j \times j$ symmetric covariance matrix of the observed variable residuals. To estimate the factor model, it is necessary to impose constraints on certain of the model parameters. It is common to restrict $\boldsymbol{\Theta}$ to be a diagonal matrix such that all residual covariances are fixed to zero, and only the residual variances on the diagonal are estimated as model parameters.

$$\boldsymbol{\Theta} = \begin{bmatrix} \theta_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{44} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_{66} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{77} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{88} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta_{99} \end{bmatrix}$$

1.2 Exploratory and Confirmatory Factor Analysis

In exploratory factor analysis (EFA), each of the j indicators has an estimated factor loading λ_{jk} on each of the k factors, but in confirmatory factor analysis (CFA), the factor loadings and/or factor variances are sometimes constrained to define the structure of the model and obtain unique estimates of the remaining parameters. For example, one common pattern of constraints is to fix the factor variances to 1, and to fix certain factor loadings to 0.

$$\boldsymbol{\Psi} = \begin{bmatrix} 1.0 & \psi_{12} & \psi_{13} \\ \psi_{21} & 1.0 & \psi_{23} \\ \psi_{31} & \psi_{32} & 1.0 \end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & 0 & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & \lambda_{52} & 0 \\ 0 & \lambda_{62} & 0 \\ 0 & 0 & \lambda_{73} \\ 0 & 0 & \lambda_{83} \\ 0 & 0 & \lambda_{93} \end{bmatrix}$$

1.3 The Independent Clusters Correlated Factors CFA Model

The first three indicators (1–3) load on the first factor, the next three (4–6) on the second factor, and the last three (7–9) on the third factor (see Figure 3.1). This is often described as a confirmatory independent cluster structure, a special form of so-called simple structure, where each indicator loads on exactly one factor. Because the correlations among the factors are freely estimated, this model is also known as the correlated factors CFA. The higher these correlations are, the more they exhibit a pattern that might be called mutual convergent validity. On the other hand, the lower these correlations are, the more they exhibit what might be called mutual discriminant validity.

If the correlations do not differ from 1.0, a single factor model could account for the data just as well as three factors, and this is consistent with the notion that there is only one reliable construct (see Figure 3.2).

If the correlations do not differ from 0.0, an orthogonal factors model could account for the data just as well as correlated factors, and this is consistent with the notion that the three separate constructs are fully independent (see Figure 3.3).

These latter two extreme possibilities are seldom observed empirically. More commonly, modest to moderate inter-factor correlations are expected and obtained.

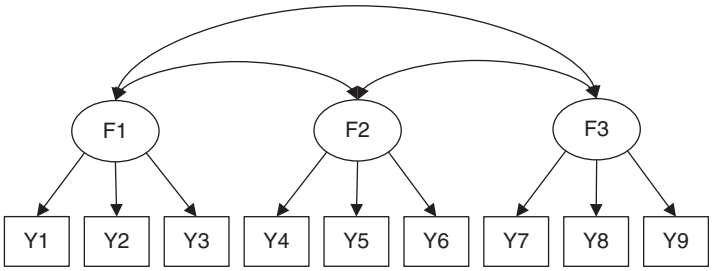


Figure 3.1 Correlated factors confirmatory factor analysis (CFA). In this and following path diagrams, indicator residual variances have been omitted for simplicity.

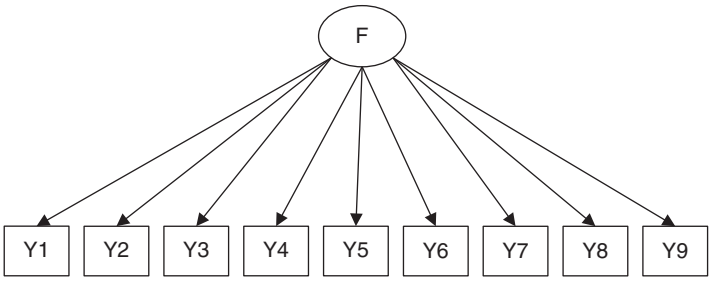


Figure 3.2 Unidimensional model confirmatory/exploratory factor analysis (CFA/EFA).

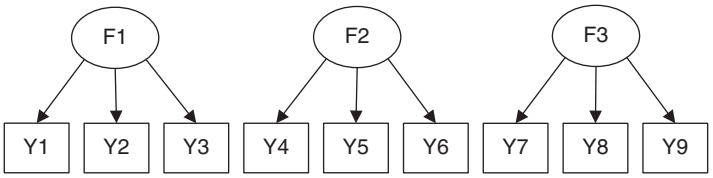


Figure 3.3 Uncorrelated (orthogonal) factors confirmatory factor analysis (CFA) model.

1.4 The Correlated Factors Exploratory Factor Analysis (EFA) Model

The correlated factors CFA is a specially restricted form of a more general factor analysis model allowing all of the items to load on all of the factors. This correlated factors exploratory factor analysis (EFA) model (see Figure 3.4) ideally approximates a simple structure with large and well-defined primary loadings, where the remaining so-called secondary loadings are generally negligible. One way to describe the relationship between the correlated factors CFA and EFA is that the CFA constrains the secondary factor loadings of the EFA to be (exactly) zero, and asserts that the CFA will nevertheless fit the data as well as the EFA, despite these constraints.

1.5 Higher-Order Factor Models

Theorists and researchers are sometimes interested in multidimensional or multifaceted constructs that exhibit both commonality and distinctiveness among multiple factors within a certain measurement domain. The correlated factors model with an independent cluster structure in Figure 3.1 can be compared with the related higher-order model in Figure 3.5. Briefly, the higher order approach models the associations among the first-order factors (F1–F3) in terms of a common second order construct or factor (H). One might say that the second order factor is to the first order factors as the first order factors are to the observed indicators. One might also say that the second order factor affects the indicators only indirectly through the first-order factors.

The idea here is that the variances and covariances of the first-order factors from the correlated factors model are a function of the higher order factor and the residuals or disturbances of the first-order factors. In the present case with just three first-order factors, the second-order model and the

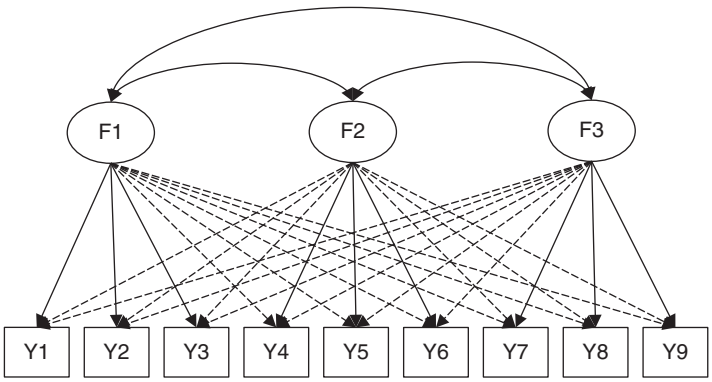


Figure 3.4 Correlated factors exploratory factor analysis (EFA). Solid path arrows represent expected primary factor loadings and dashed path arrows represent secondary factor loadings or so-called cross-loadings.

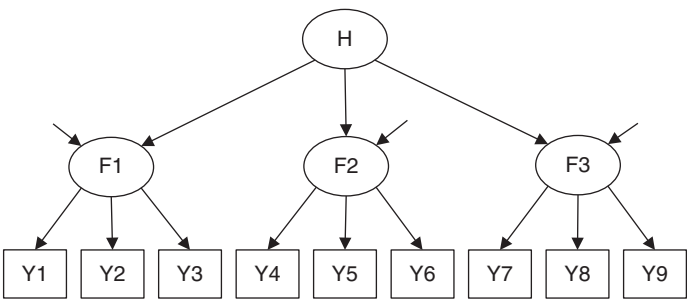


Figure 3.5 Higher-order model.

correlated factors model necessarily fit the data identically. The relative fit of the second order model in comparison with the corresponding correlated factors model can only be assessed if there are four or more first-order factors. If there were just two first-order factors, it would be necessary to impose additional constraints to be able to identify and estimate the higher order model.

We will later discuss EFA counterparts of the higher-order and other models to be presented hereafter, but will focus on the CFA versions of these models through the next few sections. Sometimes certain of the observed indicators are best conceptualized as direct indicators of the more general higher-order factor, rather than as indicators of one of the first-order factors (see Figure 3.6).

1.6 Bifactor Models

The bifactor model takes a different approach to the shared and unique components of multidimensional or multifaceted constructs. (Bornovalova, et al., 2020; Cho, 2016; Markon, 2019). All of the indicators are directly affected by both a general factor (G) as well as by additional group or specific factors (S1–S3) that are common to subsets of the indicators. The general factor is constrained to be uncorrelated with the specific factors. In general, the specific factors need not be mutually orthogonal, though we have represented them as such in Figure 3.7. In the complete independent cluster bifactor model, each indicator is influenced by the general factor, a specific factor, and an indicator-specific residual, although we have omitted these residuals from Figures 3.1–3.8 for simplicity of presentation. If each indicator is influenced by the general factor and exactly one of the specific factors, we have what is sometimes called the complete symmetric bifactor model. Schematically, the symmetric bifactor model looks like a combination of a single (general) factor model (Figure 3.2) and an orthogonal (specific) factors model (Figure 3.3).

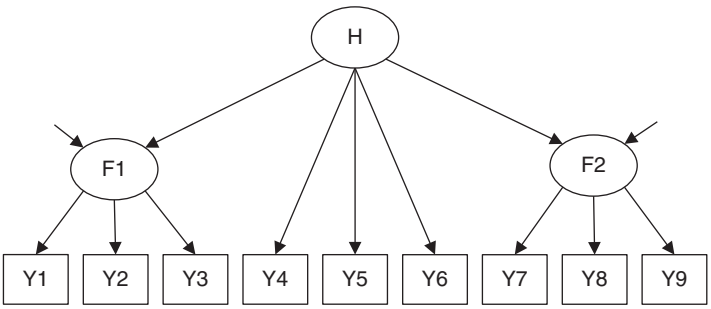


Figure 3.6 Higher-order model with direct indicators of the higher order factor.

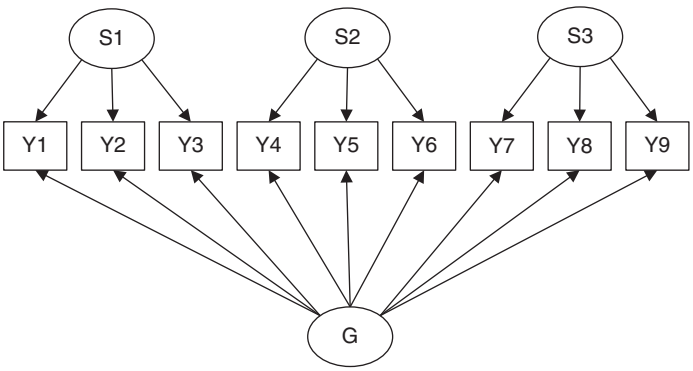


Figure 3.7 Complete (symmetric) bifactor model.

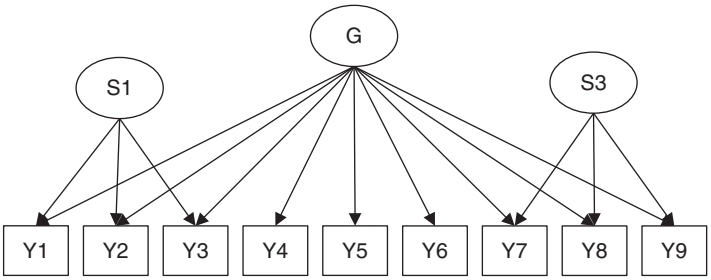


Figure 3.8 The (S-1) bifactor model with direct indicators of the general factor.

A potential problem with the symmetric bifactor model is the possibility of over-extracting factors and overfitting the data. If three dimensions or factors are sufficient to model and fit the data, but the symmetric bifactor model superfluously employs four factors (one general factor and three specific factors), there can be problems of empirical under-identification, sometimes signaled by negative variance estimates or very low factor loadings for one or more of the specific factors. By definition, variances cannot be negative, and negative variance estimates are inadmissible as model parameters in CFA. This possibility of specific "factor collapse" in the symmetric bifactor model can affect the model's interpretation. Although a corresponding higher order model also uses four factors, the higher order fourth factor is defined in terms of the variances and covariances of the first-order factors, and anomalies such as negative residual variance estimates for the first-order factors are possible but are usually less frequent.

Another useful variation of the bifactor model is the S minus 1 (S-1) bifactor model[1] shown in Figure 3.8 where *S* is the total number of specific factors that would be needed to link each indicator with an associated specific factor, as in the complete symmetric bifactor model. In the S-1 bifactor model, several of the items (in this case, items 4–6) are modeled as direct indicators of the general factor and are uninfluenced by either of the two remaining specific factors. This is a convenient special case of the bifactor model that can be less likely to suffer from convergence or estimation problems, but it requires the direct indicators of the general factor to be conceptually and empirically free from other shared influence.

The factor covariance and factor loading matrices of the S-1 bifactor model can be given as follows.

$$\boldsymbol{\Psi} = \begin{bmatrix} 1.0 & 0 & 0 \\ 0 & 1.0 & 0 \\ 0 & 0 & 1.0 \end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & 0 \\ \lambda_{21} & \lambda_{22} & 0 \\ \lambda_{31} & \lambda_{32} & 0 \\ \lambda_{41} & 0 & 0 \\ \lambda_{51} & 0 & 0 \\ \lambda_{61} & 0 & 0 \\ \lambda_{71} & 0 & \lambda_{73} \\ \lambda_{81} & 0 & \lambda_{83} \\ \lambda_{91} & 0 & \lambda_{93} \end{bmatrix}$$

1 When we speak of an "S-1 Bifactor model" it is important to remember that the "-" between the S and the 1 is a minus sign and not a hyphen. The ubiquity of this usage by factor analysts is illustrated by the wording in the title of the reference by Eid (2020) at the end of this computational guide, that speaks of "the Bifactor S-1 model". That four word phrase is pronounced aloud as "the Bifactor S-minus-one model.

In the S-1 bifactor model (see Figure 3.8), all of the indicators load on the general factor G, where indicators y4 through y6 are direct measures of (only) the general factor. Items y1 through y3 also load on one of the specific factors, and items y7 through y9 also load on another specific factor. The S-1 bifactor model sometimes arises from the collapse of a specific factor when estimating the symmetric bifactor model. In this situation, consequent model respecification by dropping the vanished specific factor produces a model where the general factor is now defined directly by the indicators of the omitted specific factor. The S-1 bifactor model with indicators designed to directly assess the general construct is preferred. An S-1 bifactor model resulting from failed estimation of a symmetric bifactor model implicitly requires an interpretive shift which treats the general factor as a reference facet corresponding to the omitted specific factor. Burns et al. (2020) and Eid (2020) advocate the use of an S-1 bifactor model where the desired reference facet is deliberately chosen.

1.7 Data Example

We will use the classic Holzinger & Swineford (1939) dataset with data on 26 mental tests administered to 301 students from two schools. Because two of the tests were only administered in one of the schools, we work with the remaining 24 tests to illustrate the correlated factor, higher-order factor, and bifactor models. We will use the HS.data[2] version from the MBESS R package that contains all 26 variables, and start with just one of the schools, designated the Grant-White school with 145 students. HS.data contains general measures of mental ability as well as measures of specific narrower cognitive functioning constructs. Although we will concentrate primarily on confirmatory factor analyses (CFA), we begin with the Mplus commands using maximum likelihood estimation to perform a four factor exploratory factor analysis (EFA) for the first 19 variables in the dataset measuring four specific mental abilities.

We also briefly distinguish here between common factor analysis and principal components analysis, which is often the default factor extraction method in general purpose statistical software such as SAS or SPSS. The principal components method attempts to model or account for all of the variance in a set of variables. In cases where nearly all of the variance can be explained by a small number of principal components relative to the total number of variables, the remainder is considered ignorable and further work is focused on these extracted components. On the other hand, common factor analysis tries to account for the variance that is common or shared among the variables, distinguishing this from variation that is specific to any given observed variable. Principal components analysis is based on a direct mathematical decomposition of the eigenvalues and eigenvectors of the variable covariance or correlation matrix, whereas common factor analysis uses iterative methods to obtain mathematically optimal estimates of its model parameters.

---

2 Steps to create the Holzinger & Swineford dataset, hsdata.dat, to beused in the ten Mplus code boxes.

1) At the end of this computational guide there are two entries in the final section entitled **Online Datasets and Services**. The first (MBESS) contains the documentation for the dataset, and the second (OpenMx) contains the .csv file for the dataset itself. (https://rdrr.io/cran/MBESS/man/HS.data.html)
2) When the second URL is entered into a browser and activated, it automatically and immediately deposits a .csv file entitled HS.csv into the downloads folder of the computer. This is a "comma separated values" file (csv).
3) This csv file is a 302×34 matrix of data. That is, it contains 302 rows (a single row for variable labels + 156 rows of Pasteur student data + 145 rows of Grant-White student data) and 34 columns (8 student categories + 26 measured test variables). However, Mplus doesn't accept text in its input data files. We must therefore create a copy the HS.cvs file and name this copy file hsdata.dat. This will be our working data. We delete the labels, that is the first row, from this file and store it in a folder where we plan to run the Mplus data analyses and store the results.

For each of the Mplus code segments, such as the first one that starts off Section 2, Subsection 2.1, to run the code one only has to place the code in the same folder as the input data. For example we might create a folder named Analyses of 12 FA Models, place the newly created hsdata.dat file in it and then run the first Mplus code in this computational guide.

With both principal components and common factor analysis, it is typical to present or plot the correlation matrix eigenvalues, accompanied by methods such as scree plots or parallel analysis to help determine the optimal number of components or factors to extract. For the Holzinger–Swineford Grant-White data, the first six eigenvalues of the correlation matrix were 6.304, 1.947, 1.527, 1.488, 0.940, and 0.875, indicating potential optimality of a four-factor solution for the 19 variables.

Though mathematically optimal, these initially extracted factors or components are generally difficult to interpret and can be complicated to explain or communicate. It is common to transform these initial results to more readily interpretable solutions using rotation algorithms designed to approximate "simple structure." Although these algorithms and the criteria they use can be quite complicated, we will consider two simplified definitions of simple structure: the (i) independent cluster structure, with each variable having a large "primary" loading on (just) one of the several factors, and (ii) the bifactor structure, with each variable loading on a general factor, and also on just one additional secondary or specific factor. In exploratory factor analysis, there are also two general factor rotation methods: those that permit the factors to be correlated with each other (oblique rotations), and those that require the factor intercorrelations to be zero (orthogonal rotations).

2 Correlated Factors Model

2.1 The Correlated Factors Exploratory Factor Analysis (EFA) Model

Here we present the Mplus syntax [EFA19_3-5.inp] to estimate a four-factor exploratory common factor analysis with data from the Grant-White school based on the first 19 cognitive variables. The default estimation method is maximum likelihood, and the default geomin rotation approximates the independent cluster approach to simple structure.

```
TITLE: Four correlated factors EFA with simple structure rotation
! Define data and select observations
DATA: FILE IS hsdata.dat;                          ! Data file
VARIABLE: NAMES ARE                                ! Variable names
 id Gender grade agey agem school                  ! Background
 visual cubes paper flags                          ! Spatial
 general paragrap sentence wordc wordm             ! Verbal
 addition code counting straight                   ! Speeded
 wordr numberr figurer object numberf figurew      ! Memory
 deduct numeric problemr series arithmet           ! Ability
 paperrev flagssub;                                ! Replacement
MISSING ARE ALL (-999);                            ! Missing values
USEOBSERVATIONS ARE (school==1);                   ! Select Grant-White
                                                     school

! Exploratory factor analysis (EFA)
USEVARIABLES ARE visual-figurew;                   ! Select first
                                                     19 variables
ANALYSIS: TYPE IS EFA 3 5;                         ! 3-5 Factors
```

In addition to listed or plotted eigenvalues, maximum likelihood exploratory factor analysis provides Chi-square model comparison tests for evaluating different numbers of factors. In this case we are examining and comparing the model fit of 3, 4, and 5 factor models for the first 19 variables in the dataset.

```
SUMMARY OF MODEL FIT INFORMATION
                  Number of                    Degrees of
     Model        Parameters      Chi-Square    Freedom     P-Value
     3-factor         92            188.747       117       0.0000
     4-factor        108            110.304       101       0.2476
     5-factor        123             82.692        86       0.5810

                                               Degrees of
     Models Compared              Chi-Square    Freedom     P-Value
     3-factor against 4-factor        78.442        16       0.0000
     4-factor against 5-factor        27.612        15       0.0241
```

From the model Chi-square values, the three-factor model shows significant model misfit, but the four and five factor models fit the data well by this criterion. The model fit improvement from three to four factors is highly significant ($p < 0.0001$), but the improvement from four to five factors has a much larger p-value ($p = 0.0241$). Based on these results, a four-factor EFA model would seem justifiable. Other model fit information is also available for the 3, 4, and 5 factor models (see Table 3.1). The three-factor model shows acceptable fit by most criteria, but the four-factor model fits much more closely by these measures. The five-factor model gives some evidence of overfitting based on the BIC, CFI, and TLI.

These measures of model fit address the notion of dimensionality in the specific case of exploratory analysis where indicators can load on all factors. In our case, three factors seem to be two few, and five factors seem to be too many, leaving our preference for four factors or dimensions.

The factor loadings for the four-factor analysis are given in Table 3.2. These factor loadings are standardized regression coefficients, sometimes called factor pattern coefficients, from a regression where the item is treated as the dependent variable, and the factors are used as the predictors.

Many general purpose statistical software programs such as SAS, SPSS, or Stata allow the user to suppress factor loadings whose absolute values are below a certain threshold. For example, if the threshold were set at 0.40 in Table 3.1, only the bolded loadings would be displayed and the rest would be blanked out.

2.2 Communalities and Exploratory Structural Equation Modeling in Mplus

Table 3.2 also includes the communalities for each variable or indicator, which are defined as the proportion of the variance in the indicator that can be jointly explained by all of the factors, or the item R-squared value. The pattern of the loadings across the factors and the communalities

Table 3.1 Additional model fit information for three, four, and five factor models with 19 variables.

	Number of factors		
	Three	**Four**	**Five**
Parameters	92	108	123
AIC	18257	18210	18213
BIC	18531	18532	18579
RMSEA	0.065	0.025	0.000
CFI	0.928	0.991	1.000
TLI	0.895	0.984	1.000
SRMR	0.058	0.034	0.028

Table 3.2 Factor loadings from a four-factor exploratory factor analysis with oblique (geomin) rotation for 19 mental tests.

	SPATIAL	VERBAL	SPEEDED	MEMORY	COMMUN
VISUAL	**0.628**	0.065	0.091	0.085	**0.535**
CUBES	**0.485**	0.050	0.007	−0.003	**0.258**
PAPER	**0.406**	0.107	0.084	0.083	**0.288**
FLAGS	**0.579**	0.160	0.013	0.026	**0.451**
GENERAL	0.042	**0.752**	0.126	−0.043	**0.656**
PARAGRAP	0.021	**0.804**	−0.056	0.098	**0.694**
SENTENCE	−0.039	**0.844**	0.085	−0.057	**0.713**
WORDC	0.094	**0.556**	0.197	0.019	**0.507**
WORDM	0.004	**0.852**	−0.074	0.069	**0.730**
ADDITION	−0.302	0.029	**0.824**	0.078	**0.640**
CODE	0.012	0.050	**0.479**	0.279	**0.447**
COUNTING	0.045	−0.159	**0.826**	−0.014	**0.623**
STRAIGHT	0.346	0.043	**0.570**	−0.055	**0.589**
WORDR	−0.024	0.117	−0.020	**0.523**	**0.316**
NUMBERR	0.069	0.021	−0.026	**0.515**	**0.290**
FIGURER	0.354	−0.033	−0.077	**0.515**	**0.440**
OBJECT	−0.195	0.045	0.154	**0.685**	**0.530**
NUMBERF	0.225	−0.127	0.246	**0.450**	**0.427**
FIGUREW	0.069	0.099	0.058	**0.365**	**0.223**

The bolded loadings indicate that the several bolded variables are best representing the factor that is named at the top of their column.

are often used to assess the role and contribution of the item to the analysis. Mplus' exploratory factor analysis (EFA) does not provide the communalities. However, one can use Mplus' exploratory structural equation modeling (ESEM) approach [ESEM19.inp] to obtain the communalities. This specification reproduces exactly the same factor loadings as the traditional EFA, but allows the request of standardized results, including the communalities.

```
...
! Exploratory structural equation modeling (ESEM) with four factors
USEVARIABLES ARE visual-figurew;      ! Select first 19 variables
MODEL: F1-F4 BY visual-figurew (*1);  ! Specify the model
OUTPUT: STDYX;                        ! Standardized results
```

In the EFA and ESEM analyses, we used Mplus' default oblique rotation (geomin) which permits the estimated factors to be correlated with each other. Mplus and other software also offer other oblique rotations such as promax, direct oblimin, etc. In the present data, the inter-factor correlations are fairly uniform, averaging 0.366 (see Table 3.3).

This correlated common factors model is sometimes described as a multidimensional or multi-unidimensional model, reflecting multiple theoretically distinguishable, but potentially empirically correlated, constructs or dimensions.

Table 3.3 Inter-factor correlations from a four factor exploratory factor analysis with an oblique rotation.

	SPATIAL	VERBAL	SPEEDED	MEMORY
SPATIAL	1.000			
VERBAL	0.378	1.000		
SPEEDED	0.372	0.386	1.000	
MEMORY	0.307	0.380	0.375	1.000

2.3 The Correlated Factors Confirmatory Factor Analysis (CFA) Model

Based on sufficiently strong substantive theory or as a confirmatory follow-on to an EFA such as that shown above, researchers sometimes propose a model where only the loadings for certain designated indicators of a given factor are estimated, and all other loadings are fixed to zero. The Mplus syntax for the four-factor correlated factors CFA [CFA19.inp] is given below.

```
...
! Confirmatory factor analysis
USEVARIABLES ARE visual-figurew; ! Select first 19 variables
MODEL: spatial BY visual cubes paper flags;
 verbal  BY general paragrap sentence wordc wordm;
 speed   BY addition code counting straight;
 memory  BY wordr numberr figurer object numberf figurew;
OUTPUT: STDYX; TECH4;
```

The TECH4 keyword on the OUTPUT: command produces the correlation matrix of the latent variables. Standard errors and significance tests for these correlations are available elsewhere in the output. Compared to the EFA correlations given in Table 3.3, the corresponding CFA correlations in Table 3.4 are noticeably larger. This is a common result and has implications for the nature of the latent variables that are intended to be used and analyzed in other analyses.

3 Higher-Order Factor Models

3.1 The Higher-Order Factor CFA

Where there are moderate to substantial correlations among the factors, one can consider a higher-order structure, where these associations among the factors are believed to be explained by a common second-order construct. Here we identify and scale the four first order latent variables (*spatial*, *verbal*, *speed*, and *memory*) and the second order latent variable *mental* using the usual Mplus default method which automatically fixes the first loading for each factor to 1.0 [CFA19_HO.inp]. Because we are

Table 3.4 Inter-factor correlations from a four-factor exploratory factor analysis with oblique rotation.

	SPATIAL	VERBAL	SPEEDED	MEMORY
SPATIAL	1.000			
VERBAL	0.594	1.000		
SPEEDED	0.583	0.467	1.000	
MEMORY	0.635	0.496	0.605	1.000

interested in the standardized loadings, we use the STDYX keyword on the OUTPUT: command to obtain the completely standardized factor loadings and other estimates.

```
...
! Confirmatory factor analysis: Higher-order model
USEVARIABLES ARE visual-figurew;              ! Select first 19 variables
MODEL: spatial BY visual cubes paper flags;
 verbal  BY general paragrap sentence wordc wordm;
 speed   BY addition code counting straight;
 memory  BY wordr numberr figurer object numberf figurew;
 mental  BY spatial verbal speed memory;
 mental@1;
OUTPUT: STDYX; TECH4;
```

The first- and second-order standardized factor loadings for the second order model are given in Table 3.5, along with the associated communalities. In this model, again, all of the non-target factor loadings are fixed to be exactly zero.

Table 3.5 First- and second-order factor loadings, and communalities, for the second order factor model with 19 mental tests.

	SPATIAL	VERBAL	SPEED	MEMORY	MENTAL	COMMUN
VISUAL	0.696					0.485
CUBES	0.478					0.228
PAPER	0.543					0.295
FLAGS	0.670					0.449
GENERAL		0.807				0.651
PARAGRAP		0.820				0.672
SENTENCE		0.837				0.701
WORDC		0.696				0.484
WORDM		0.843				0.710
ADDITION			0.643			0.414
CODE			0.690			0.477
COUNTING			0.698			0.488
STRAIGHT			0.759			0.575
WORDR				0.519		0.270
NUMBERR				0.521		0.272
FIGURER				0.603		0.363
OBJECT				0.625		0.390
NUMBERF				0.644		0.415
FIGUREW				0.484		0.234
SPATIAL					**0.839**	0.704
VERBAL					**0.668**	0.446
SPEED					**0.727**	0.528
MEMORY					**0.776**	0.602

This is a higher order factor model with four first order factors that are the last four variables listed and bolded on the left combined into a second order factor. They are combined into a second order factor named MENTAL, and the bolded loadings show the relationship between the four first order factor and this MENTAL second order factor.

3.2 The Higher Order CFA Model with Direct Indicators of the Higher Order Factor

In addition to measures of specific narrower cognitive abilities, the Holzinger–Swineford data also contains a number of direct measures of more general mental functioning. The model which also includes these measures has 24 total indicators and can be implemented in the following Mplus syntax [CFA24_HODI.inp] (Table 3.6).

Table 3.6 Higher-order CFA for 24 tests of mental ability with direct indicators of the higher order factor.

	SPATIAL	VERBAL	SPEED	MEMORY	MENTAL	COMMUN
VISUAL	0.695					0.483
CUBES	0.503					0.253
PAPER	0.503					0.253
FLAGS	0.688					0.473
GENERAL		0.806				0.650
PARAGRAP		0.816				0.666
SENTENCE		0.836				0.699
WORDC		0.699				0.489
WORDM		0.846				0.715
ADDITION			0.672			0.451
CODE			0.677			0.459
COUNTING			0.720			0.518
STRAIGHT			0.734			0.539
WORDR				0.524		0.275
NUMBERR				0.516		0.266
FIGURER				0.606		0.367
OBJECT				0.626		0.392
NUMBERF				0.629		0.396
FIGUREW				0.501		0.251
DEDUCT					0.653	0.427
NUMERIC					0.639	0.408
PROBLEMR					0.653	0.426
SERIES					0.732	0.535
ARITHMET					0.636	0.405
SPATIAL					**0.804**	0.646
VERBAL					**0.733**	0.537
SPEED					**0.685**	0.469
MEMORY					**0.762**	0.581

This Table 3.6 is essentially the same as the higher order factor model presented in Table 3.5, but with five more primary variables in addition to the same 19 primary variables used in Table 3.5 that were extracted to form the four first order factors (spatial, verbal, speed, and memory). These five new primary variables, when combined with the four first order factors become direct indicators of the higher order factor.

```
...
! Confirmatory factor analysis: Higher-order model with direct indicators
USEVARIABLES ARE visual-arithmet;              ! Select all 24 variables
MODEL: spatial BY visual cubes paper flags;
 verbal  BY general paragrap sentence wordc wordm;
 speed   BY addition code counting straight;
 memory  BY wordr numberr figurer object numberf figurew;
 mental  BY spatial* verbal speed memory
            deduct* numeric problemr series arithmet;
 mental@1;
OUTPUT: STDYX; TECH4;
```

3.3 The Higher Order EFA/ESEM Model with Direct Indicators of the Higher Order Factor

Because the Holzinger–Swineford data also contain a set of indicators of general mental ability, we also estimated a higher order EFA model for all 24 mental tests, with direct assessment of the higher order factor by the five additional general mental ability tests (*deduct*, *numeric*, *problem*, *series*, and *arithmet*) [ESEM24_HODI.inp] (Tables 3.7 and 3.8).

Table 3.7 Higher-order model with a directly assessed higher order factor for 24 mental tests.

	SPATIAL	VERBAL	SPEED	MEMORY	MENTAL	COMMUN
VISUAL	**0.655**	0.008	0.062	0.075		0.537
CUBES	**0.541**	0.003	−0.013	−0.016		0.280
PAPER	**0.390**	0.082	0.060	0.067		0.262
FLAGS	**0.630**	0.113	−0.019	0.002		0.469
GENERAL	0.036	**0.791**	0.082	−0.088		0.655
PARAGRAP	−0.002	**0.847**	−0.123	0.073		0.688
SENTENCE	−0.050	**0.896**	0.037	−0.099		0.712
WORDC	0.084	**0.573**	0.163	−0.007		0.510
WORDM	−0.004	**0.902**	−0.142	0.042		0.738
ADDITION	−0.365	0.025	**0.864**	0.081		0.641
CODE	−0.061	0.033	**0.468**	0.295		0.431
COUNTING	0.020	−0.201	**0.888**	−0.032		0.643
STRAIGHT	0.322	0.011	**0.575**	−0.080		0.556
WORDR	−0.096	0.106	−0.071	**0.580**		0.318
NUMBERR	0.012	−0.005	−0.068	**0.563**		0.285
FIGURER	0.322	−0.087	−0.137	**0.575**		0.450
OBJECT	−0.296	0.031	0.112	**0.752**		0.521

(*Continued*)

Table 3.7 (Continued)

	SPATIAL	VERBAL	SPEED	MEMORY	MENTAL	COMMUN
NUMBERF	0.182	−0.182	0.223	**0.488**		0.426
FIGUREW	0.033	0.079	0.015	**0.413**		0.236
DEDUCT					**0.655**	0.428
NUMERIC					**0.629**	0.396
PROBLEMR					**0.651**	0.424
SERIES					**0.729**	0.532
ARITHMET					**0.641**	0.411
SPATIAL					**0.704**	0.496
VERBAL					**0.767**	0.589
SPEED					**0.686**	0.471
MEMORY					**0.724**	0.524

Table 3.8 Standardized residual variances, covariances, and correlations among the first order factors when controlling for the higher order factor.

	SPATIAL	VERBAL	SPEED	MEMORY
SPATIAL	0.504	−0.211	−0.010	−0.086
VERBAL	−0.096	0.411	−0.146	−0.199
SPEED	−0.005	−0.068	0.529	0.060
MEMORY	−0.042	−0.088	0.030	0.476

Note: The standardized residual variances are on the diagonal, the covariances are below the diagonal, and the correlations are above the diagonal in Table 3.8.

```
...
! Higher order exploratory structural equation modeling direct indicators
USEVARIABLES ARE visual-arithmet;  ! Use all 24 tests
MODEL: F1-F4 BY visual-figurew (*1); ! Four exploratory first order factors
 G BY F1-F4* deduct-arithmet*;      ! Higher order factor direct indicators
 G@1;                               ! Fix general factor variance to 1
OUTPUT: STDYX; TECH4;               ! Standardized estimates;
```

4 Bifactor Models

4.1 The Symmetric or Complete Bifactor CFA

In this model, every item loads on exactly one specific factor and on the general factor. This yields an independent cluster structure for the specific factors. This is the model shown in Figure 3.3. Compared to a correlated factor model with K factors, this model has K + 1 total factors (the K specific factors plus the general factor). The Mplus syntax to perform this analysis is listed below [CFA19_BI.inp].

```
...
! Confirmatory factor analysis: Symmetric bifactor model
! Fails to converge
USEVARIABLES ARE visual-figurew;  ! Select the first 19 tests
ANALYSIS: MODEL IS NOCOVARIANCES; ! Omit factor covariances
MODEL: G BY visual-figurew;       ! General factor
 spatial BY visual-flags;         ! Specific factors
 verbal  BY general-wordm;        !
 speed   BY addition-straight;    !
 memory  BY wordr-figurew;        !
OUTPUT: STDYX; TECH4; ! Standardized estimates & latent correlations
```

With the Holzinger–Swineford data, this produces a phenomenon sometimes known as "factor collapse" with bifactor models where the loadings for one of the specific factors, or the specific factor variance, become vanishingly small. In this case, Mplus failed to converge, but produced the following intermediate output for the factor variances.

```
Variances
    G                 23.667
    SPATIAL            0.001
    VERBAL            50.937
    SPEED            301.018
    MEMORY            20.777
```

Although these values are generally not fully trustworthy, there is evidence that the SPATIAL specific factor does not provide variance beyond that due to the general factor. This is somewhat understandable, because five factors are now being estimated, whereas there were only four in the corresponding correlated factors model. The added general factor has direct effects on each of the items, and this can result in possible over-factoring in such cases. One possible remedy to this is to omit the SPATIAL general factor and re-estimate the model. In this context, this is equivalent to treating SPATIAL as a "reference construct" that is identical to the estimated general factor which affects all of the items. In the Mplus syntax, this can be accomplished by "commenting out" the SPATIAL factor by putting an exclamation point in front of it in the above MODEL: syntax.

```
...
! spatial BY visual-flags;
...
```

4.2 The Bifactor CFA with a Reference Factor

Because the traditional symmetric bifactor model led to collapse of the SPATIAL specific factor, the model was re-estimated excluding the specific factor for these items. The following output gives the factor loadings for the resulting model. This model includes the general factor, which is identified with the original SPATIAL specific factor, and the three other specific factors (VERBAL, SPEED, and MEMORY) (Table 3.9).

Table 3.9 Bifactor CFA for 19 items treating SPATIAL as a (general) reference factor.

	GENERAL/SPATIAL	VERBAL	SPEED	MEMORY
VISUAL	0.709			
CUBES	0.463			
PAPER	0.552			
FLAGS	0.644			
GENERAL	0.517	0.615		
PARAGRAP	0.487	0.664		
SENTENCE	0.453	0.713		
WORDC	0.534	0.459		
WORDM	0.468	0.707		
ADDITION	0.259		0.737	
CODE	0.481		0.439	
COUNTING	0.401		0.641	
STRAIGHT	0.649		0.386	
WORDR	0.312			0.428
NUMBERR	0.319			0.418
FIGURER	0.519			0.311
OBJECT	0.286			0.684
NUMBERF	0.507			0.385
FIGUREW	0.357			0.307

Factor collapse is not inevitable in complete symmetric bifactor analyses, but is a relatively common occurrence. If there are direct indicators of the general factor that do not load on any of the specific factors, one can often simultaneously estimate the general factor and all of the other specific factors.

4.3 The Bifactor CFA Model with Direct Indicators of the General Factor

Because the Holzinger–Swineford data also contain a set of indicators of general mental ability, we also estimated a bifactor CFA containing all 24 mental tests [CFA24_BI.inp]. This corresponds with the S-1 bifactor model discussed earlier where the general mental ability factor is directly assessed by the five additional general tests (Table 3.10).

```
...
USEVARIABLES ARE visual-arithmet; ! Select all 24 tests
ANALYSIS: MODEL IS NOCOVARIANCES; ! Omit factor covariances
MODEL: G BY visual-arithmet;      ! General factor
 spatial BY visual-flags;         ! Specific factors
 verbal  BY general-wordm;        !
 speed   BY addition-straight;    !
 memory  BY wordr-figurew;        !
OUTPUT: STDYX; TECH4; ! Standardized estimates & latent correlations
```

Table 3.10 Orthogonal bifactor CFA for 24 tests of mental ability.

	MENTAL	SPATIAL	VERBAL	SPEED	MEMORY	COMMUN
VISUAL	0.606	0.235				0.422
CUBES	0.400	0.337				0.273
PAPER	0.422	0.238				0.235
FLAGS	0.554	0.546				0.604
GENERAL	0.597		0.532			0.640
PARAGRAP	0.574		0.592			0.680
SENTENCE	0.573		0.619			0.711
WORDC	0.617		0.349			0.502
WORDM	0.599		0.600			0.719
ADDITION	0.402			0.610		0.535
CODE	0.507			0.395		0.413
COUNTING	0.448			0.635		0.604
STRAIGHT	0.578			0.400		0.494
WORDR	0.378				0.401	0.303
NUMBERR	0.349				0.425	0.303
FIGURER	0.513				0.292	0.348
OBJECT	0.410				0.557	0.478
NUMBERF	0.510				0.335	0.373
FIGUREW	0.440				0.207	0.237
DEDUCT	0.639					0.409
NUMERIC	0.645					0.416
PROBLEMR	0.648					0.419
SERIES	0.726					0.527
ARITHMET	0.620					0.384

The last five items directly measure the overall mental ability factor and therefore only load on the general factor.

4.4 The Orthogonal Bifactor EFA/ESEM with Direct Indicators of the General Factor

As with CFA, one can estimate an orthogonal exploratory bifactor model using the following Mplus syntax to fit the model and obtain the standardized factor loadings and communalities [ESEM24_BI_DI.inp] (Table 3.11).

```
...
ANALYSIS: ROTATION IS BI-GEOMIN(ORTHOGONAL);
MODEL: G F1-F4 BY visual-arithmet (*1);
OUTPUT: STDYX;
```

Table 3.11 Orthogonal bifactor EFA for 24 specific and general tests of mental ability.

	MENTAL	SPATIAL	VERBAL	SPEED	MEMORY	COMMUN
VISUAL	**0.602**	**0.405**	−0.056	0.004	0.003	0.529
CUBES	**0.434**	**0.220**	−0.101	−0.106	−0.130	0.275
PAPER	**0.401**	**0.363**	0.067	0.117	0.097	0.320
FLAGS	**0.579**	**0.326**	−0.024	−0.103	−0.115	0.466
GENERAL	**0.612**	−0.013	**0.516**	0.030	−0.070	0.647
PARAGRAP	**0.566**	0.034	**0.611**	−0.043	0.111	0.710
SENTENCE	**0.576**	−0.024	**0.613**	0.021	−0.054	0.712
WORDC	**0.613**	0.036	**0.350**	0.070	−0.045	0.507
WORDM	**0.608**	−0.030	**0.593**	−0.144	0.007	0.743
ADDITION	**0.466**	−0.403	−0.042	**0.587**	0.001	0.726
CODE	**0.491**	0.023	0.051	**0.432**	0.321	0.534
COUNTING	**0.517**	−0.103	−0.214	**0.496**	−0.097	0.579
STRAIGHT	**0.568**	0.327	0.016	**0.512**	0.000	0.692
WORDR	**0.368**	−0.034	0.076	−0.029	**0.459**	0.354
NUMBERR	**0.375**	−0.011	−0.041	−0.088	**0.368**	0.285
FIGURER	**0.511**	0.176	−0.117	−0.111	**0.362**	0.449
OBJECT	**0.448**	−0.237	−0.015	0.020	**0.464**	0.473
NUMBERF	**0.557**	−0.031	−0.220	0.068	**0.227**	0.415
FIGUREW	**0.427**	−0.047	−0.004	−0.004	**0.227**	0.236
DEDUCT	**0.645**	−0.022	0.082	−0.223	0.004	0.473
NUMERIC	**0.643**	0.034	−0.129	0.136	−0.040	0.451
PROBLEMR	**0.635**	0.011	0.086	−0.183	0.031	0.445
SERIES	**0.729**	0.038	0.030	−0.114	−0.095	0.556
ARITHMET	**0.642**	−0.308	0.082	0.180	0.044	0.548

4.5 Model Fit of Multidimensional Factor Analysis Models

Table 3.12 contains a variety of fit measures for several selected models for the first 19 variables in the Holzinger–Swineford data, including the correlated factors CFA, the higher order factor CFA, the S-1 bifactor CFA, and the correlated factors ESEM. The S-1 bifactor CFA model is used because the symmetric bifactor model failed to converge for these data.

These models generally range from acceptably to well fitting, with the bifactor CFA and correlated factors ESEM models fitting the data more closely for these 19 tests. Table 3.13 contains fit measures for the higher order CFA, the higher order ESEM, the bifactor CFA, and the bifactor ESEM models.

In general, the ESEM models fit better than the CFA models, and the bifactor models fit better than the higher order models. ESEM models and bifactor models generally estimate substantially more parameters that may not be of substantive or theoretical relevance, and may thus overfit the data or drift toward chance aspects of the data.

Table 3.12 Correlated factors, higher order, and S-1 bifactor CFA models, and a correlated factors ESEM model.

	Correlated factors CFA	Higher order factor CFA	S-1 Bifactor CFA	Correlated factors ESEM
Variables	19	19	19	19
Parameters	63	60	72	108
AIC	18225	18252	18206	18210
BIC	18413	18430	18420	18532
SA-BIC	18214	18240	18192	18190
Chi-square	215.544	247.736	177.763	110.304
df	146	149	137	101
RMSEA	0.057	0.068	0.045	0.025
CFI	0.930	0.901	0.959	0.991
TLI	0.918	0.886	0.949	0.984
SRMR	0.063	0.113	0.064	0.031

Table 3.13 Higher order and bifactor CFA and ESEM models with direct indicators of the higher order factor or the general factor.

	Higher order CFA	Higher order ESEM	Bifactor CFA	Bifactor ESEM
Variables	24	24	24	24
Parameters	76	127	91	158
AIC	23063	23056	23068	23040
BIC	23289	23434	23339	23510
SA-BIC	23049	23032	23051	23010
Chi-square	362.165	253.403	337.632	174.816
df	248	197	233	166
RMSEA	0.056	0.044	0.056	0.019
CFI	0.916	0.959	0.923	0.994
TLI	0.907	0.942	0.909	0.989
SRMR	0.066	0.047	0.060	0.031

5 The Relationship Between the Higher Order and Bifactor Models

5.1 The Higher Order Model as a Constrained Bifactor Model

The higher order model is nested within the bifactor model. More specifically, the higher order model can be obtained by placing appropriate proportionality constraints on the factor loadings of the bifactor model. Thus, traditional model comparison techniques can be used to jointly test the viability of these constraints. This can be illustrated using the following Mplus syntax [CFA24_BI_HO.inp]:

```
...
! Impose proportionality constraints on the bifactor factor loadings
!  to obtain the fit and parameters of the higher-order model
USEVARIABLES ARE visual-arithmet;        ! Select all 24 variables
ANALYSIS: MODEL IS NOCOVARIANCES;        ! Omit factor covariances
MODEL: G BY visual-arithmet*  (g1-g24);  ! General factor
 spatial BY visual-flags      (b1-b4);   ! Specific factors
 verbal  BY general-wordm     (b5-b9);   !
 speed   BY addition-straight (b10-b13); !
 memory  BY wordr-figurew     (b14-b19); !
MODEL CONSTRAINT:                        ! Model constraints
 NEW(b1 b5 b10 b14 a1-a4);               ! New parameters
 b1=1; b5=1; b10=1; b14=1; a1=1;         ! Fix marker variable
                                           loadings to 1
 DO (1,4)   g#=a1*b#;                    ! Impose parameter
                                           constraints
 DO (5,9)   g#=a2*b#;                    !
 DO (10,13) g#=a3*b#;                    !
 DO (14,19) g#=a4*b#;                    !
OUTPUT: STDYX;                           ! Standardized estimates
```

As before, we analyze all 24 variables, omitting the factor covariances. The general and specific factor loadings are labeled, and the MODEL CONSTRAINT: command is used to impose the appropriate proportionality constraints. By default, Mplus fixes the first factor loading of each of the specific factors to 1, but when these are to be used in the ensuing proportionality constraints, it is necessary to explicitly fix the loadings of these marker variables to 1. The parameters a1–a4 are treated as place-holders for the second-order factor loadings of the specific factors on the overall construct when applying the proportionality constraints. The resulting model-implied loading estimates are computed as the product of the appropriate second order parameter and the obtained first order specific factor loadings.

The higher order model estimated by placing the necessary proportionality constraints on the bifactor model is identical in fit, prediction, and data likelihood to the directly estimated higher order model. However, this model is parameterized to yield direct estimates of the loadings of the indicators on the general factor whereas the standard higher order model yields instead the loadings of the indicators on the first order factors along with the loadings of the first order factors on the higher order factor. The loadings of the higher order CFA with direct indicators of the higher order factor, are equal to the Schmid-Leiman transformed loadings of the indicators on the general factor. Similarly, the loadings of the indicators on the specific factors are the Schmid-Leiman transformed loadings on the specific factors.

5.2 Loadings of the Indicators on the Higher-Order Factor Using the Higher Order Model

In the higher order model, the indicators are a combined function of the overall higher order factor, a unique contribution of the associated first order factor disturbance, and the indicator residual or error. In the higher order model, the Schmid-Lieman transformed loadings are not directly

estimated, but can be obtained from the indirect effects of the higher order factor on the indicators, treating the first-order factors as the mediators. This can be accomplished with the following Mplus syntax [CFA24_HODI_SL.inp].

```
...
USEVARIABLES ARE visual-arithmet; ! Select all 24 variables
MODEL:                           ! Model
 spatial BY visual-flags;        ! First order factors
 verbal  BY general-wordm;       !
 speed   BY addition-straight;   !
 memory  BY wordr-figurew;       !
 mental  BY deduct-arithmet*     ! Higher order factor
  spatial@1 verbal-memory;       !  with direct indicators

 F1 BY spatial; F2 BY verbal;    ! Replace first order residuals
 F3 BY speed; F4 BY memory;      !  with latent variables
 spatial-memory@0;               ! Fix variances to 0
 mental F1-F4 WITH F1-F4@0;      ! Fix covariances to 0

MODEL INDIRECT:
! Schmid-Leiman transformed loadings on the general factor
 visual IND mental; cubes IND mental; paper IND mental; flags IND mental;
 general IND mental; paragrap IND mental; sentence IND mental;
 wordc IND mental; wordm IND mental;
 addition IND mental; code IND mental; counting IND mental; straight IND
mental;
 wordr IND mental; numberr IND mental; figurer IND mental; object IND mental;
 numberf IND mental; figurew IND mental;
! Schmid-Leiman transformed loadings on the specific factors
 visual IND F1; cubes IND F1; paper IND F1; flags IND F1;
 general IND F2; paragrap IND F2; sentence IND F2;
 wordc IND F2; wordm IND F2;
 addition IND F3; code IND F3; counting IND F3; straight IND F3;
 wordr IND F4; numberr IND F4; figurer IND F4; object IND F4;
 numberf IND F4; figurew IND F4;
OUTPUT: STDYX;
```

After the initial commands to estimate the higher order model, the disturbances of the first order factors are replaced with latent variables and certain variances and covariances are fixed to zero. The MODEL INDIRECT: command is then used to obtain the indirect effects of the higher order factor and the latent variable replacements for the first order factor disturbances on the indicators. These estimates are included in Table 3.14 along with the corresponding results from the bifactor model with proportionality constraints. The two sets of parameters are clearly identical. Either method can be used to obtain the Schmid-Leiman transformed loadings on the general and specific factors. Note that in Table 3.14, we have listed the specific factor loadings in a single column with stub headings to indicate the relevant label for the specific factor. This differs from our usual

Table 3.14 Schmid-Leiman transformed loadings for a higher order CFA with direct indicators of the higher order factor using the constrained bifactor and higher order models.

	Constrained bifactor		Higher order		Bifactor		Extended higher order	
	Direct effects		Indirect effects		Direct effects		Total effects	
	GENERAL	SPECIFIC	GENERAL	SPECIFIC	GENERAL	SPECIFIC	GENERAL	SPECIFIC
SPATIAL								
VISUAL	0.558	0.413	0.558	0.413	0.606	0.235	0.606	0.235
CUBES	0.404	0.299	0.404	0.299	0.400	0.337	0.400	0.337
PAPER	0.404	0.299	0.404	0.299	0.422	0.238	0.422	0.238
FLAGS	0.553	0.409	0.553	0.409	0.554	0.546	0.554	0.546
VERBAL								
GENERAL	0.591	0.548	0.591	0.548	0.597	0.532	0.597	0.532
PARAGRAP	0.598	0.555	0.598	0.555	0.574	0.592	0.574	0.592
SENTENCE	0.613	0.569	0.613	0.569	0.573	0.619	0.573	0.619
WORDC	0.512	0.476	0.512	0.476	0.617	0.349	0.617	0.349
WORDM	0.620	0.575	0.620	0.575	0.599	0.600	0.599	0.600
SPEED								
ADDITION	0.460	0.489	0.460	0.489	0.402	0.610	0.402	0.610
CODE	0.464	0.494	0.464	0.494	0.507	0.395	0.507	0.395
COUNTING	0.493	0.524	0.493	0.524	0.448	0.635	0.448	0.634
STRAIGHT	0.503	0.535	0.503	0.535	0.578	0.400	0.578	0.400
MEMORY								
WORDR	0.400	0.339	0.400	0.339	0.378	0.401	0.378	0.401
NUMBERR	0.393	0.334	0.393	0.334	0.349	0.425	0.349	0.425
FIGURER	0.462	0.392	0.462	0.392	0.513	0.292	0.513	0.292
OBJECT	0.477	0.405	0.477	0.405	0.410	0.557	0.410	0.557
NUMBERF	0.480	0.407	0.480	0.407	0.510	0.335	0.510	0.335
FIGUREW	0.382	0.324	0.382	0.324	0.440	0.207	0.440	0.207
ABILITY								
DEDUCT	0.653		0.653		0.639		0.639	
NUMERIC	0.639		0.639		0.645		0.645	
PROBLEMR	0.653		0.653		0.648		0.648	
SERIES	0.732		0.732		0.726		0.726	
ARITHMET	0.636		0.636		0.620		0.620	

approach of representing the specific factors in different columns. This is feasible here because all of the indicators had loadings on just one (or no) specific factor.

The first four data columns show that the bifactor model with appropriate proportionality constraints on the general and specific factor loadings produces standardized factor loadings that are

identical to those from the higher order model. The loadings of interest here are the loadings of the indicators on the general and specific factors. In the constrained bifactor model, these are estimated as the (constrained) direct effects of the general and specific latent variables on the indicators. Estimation of these effects with the higher order model is facilitated by first replacing the first order factor disturbances with corresponding latent variables, and then estimating the indirect effects of these new specific factor latent variables and the higher order factor on the indicators using Mplus' MODEL INDIRECT: command.

5.3 The Bifactor Model as an Extended Higher Order Factor Model

Just as the higher order factor model can be fitted as a constrained bifactor model, the bifactor model can be fitted as an extended higher order factor model, and the standardized loadings of the indicators on the general factor can be obtained as the standardized total effects of the general factor on the indicators. This can be accomplished using the following Mplus syntax [CFA24_HO_BI.inp].

```
...
USEVARIABLES ARE visual-arithmet;
MODEL:
 spatial BY visual cubes paper flags;
 verbal  BY general paragrap sentence wordc wordm;
 speed   BY addition code counting straight;
 memory  BY wordr numberr figurer object numberf figurew;
 H@1; H  BY spatial* verbal speed memory
  deduct* numeric problemr series arithmet
  cubes-flags paragrap-wordm code-straight numberr-figurew;
 F1 BY spatial; F2 BY verbal; ! Replace first order residuals
 F3 BY speed; F4 BY memory;   !  with latent variables
 spatial-memory@0;            ! Fix variances to 0
 H F1-F4 WITH F1-F4@0;        ! Fix covariances to 0

MODEL INDIRECT:
! Schmid-Leiman transformed loadings on the general factor
  visual IND H; cubes IND H; paper IND H; flags IND H;
  general IND H; paragrap IND H; sentence IND H; wordc IND H; wordm IND H;
  addition IND H; code IND H; counting IND H; straight IND H;
  wordr IND H; numberr IND H; figurer IND H; object IND H;
   numberf IND H; figurew IND H;
! Schmid-Leiman transformed loadings on the specific factors
  visual IND F1; cubes IND F1; paper IND F1; flags IND F1;
  general IND F2; paragrap IND F2; sentence IND F2;
  wordc IND F2; wordm IND F2;
  addition IND F3; code IND F3; counting IND F3; straight IND F3;
  wordr IND F4; numberr IND F4; figurer IND F4; object IND F4;
  numberf IND F4; figurew IND F4;
OUTPUT: STDYX; TECH4;
```

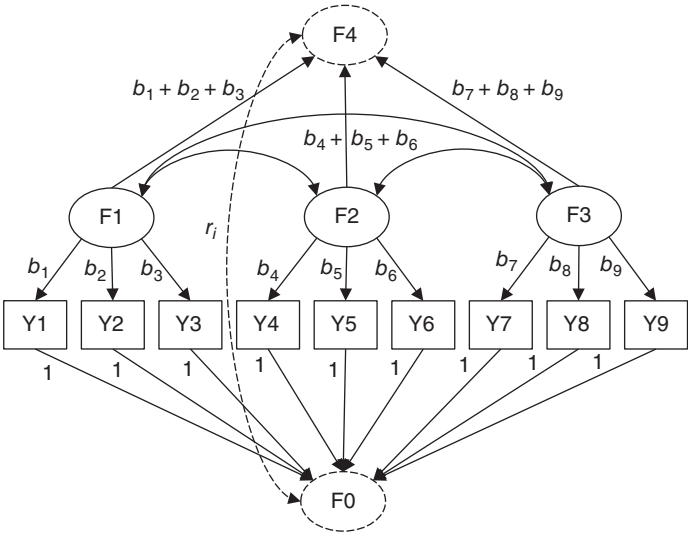


Figure 3.9 Higher order model with latent disturbances and additional direct effects of the higher-order factor on selected indicators (equivalent to the bifactor model).

This model extends the higher order factor model by adding direct effects of the higher order factor on selected indicators. Figure 3.9 gives a schematic path diagram for the model.

In addition to the usual loadings of the first order factors (F1, F2, and F3) on the higher order factor (H), H also has direct effects on indicators Y2, Y3, Y5, Y6, Y8, and Y9. Notice also that the disturbances of the first order factors have been replaced with latent variables (D1, D2, and D3). This is technically unnecessary to estimate the model, but it simplifies parameter recovery in Mplus. The first order factor disturbances are replaced with corresponding latent variables. For each of the S specific factors, each with J_s indicators, $J_s - 1$ additional direct effects are estimated where the associated standardized total effects are equal to the standardized effects of the general factor on the indicators of the bifactor model and the standardized indirect effects of the specific factor latent variables are equal to the standardized effects of the specific factors on the indicators of the bifactor model (see Table 3.14).

6 Reliability Estimation for Multidimensional Constructs

6.1 Stratified Alpha

With multidimensional data, one approach to reliability estimation for the independent clusters correlated common factors model (see Figure 3.1; Rae (2007)) can be expressed as a stratified coefficient alpha where σ_k^2 is the variance and α_k is the coefficient alpha for subscale k, and σ_T^2 is the measure's total composite variance.

$$\alpha_s = 1 - \frac{\sum_{k=1}^{K} \sigma_k^2 (1 - \alpha_k)}{\sigma_T^2}$$

Here, $\sigma_k^2 = \sum_{j_k=1}^{J_k}\sum_{j'_k=1}^{J_k}\theta_{j_k j'_k}$ is the sum of the variances and covariances of the J_k items on subscale k, $\sigma_T^2 = \sum_{j=1}^{J}\sum_{j'=1}^{J}\theta_{jj'}$ is the sum of the variances and covariances of all J items across all of the subscales, and the alpha coefficient for subscale k is:

$$\alpha_k = \frac{J_k}{J_k - 1}\left(1 - \frac{\sum_{j_k=1}^{J_k}\theta_{j_k j_k}}{\sum_{j_k=1}^{J_k}\sum_{j'_k=1}^{J_k}\theta_{j_k j'_k}}\right)$$

Stratified alpha can be obtained for the first nineteen tests in the Holzinger & Swineford data using the following Mplus syntax [SALPHA19.inp].

```
...
USEVARIABLES visual-figurew; ! Select the first 19 variables
MODEL: visual-flags (v1-v4); general-wordm   (v5-v9);   ! Subscale
                                                          variances
 addition-straight (v10-v13); wordr-figurew (v14-v19); !
 visual-flags WITH visual-flags             (c1-c6);    ! Subscale
                                                          covariances
 general-wordm WITH general-wordm          (c7-c16);   !
 addition-straight WITH addition-straight (c17-c22);   !
 wordr-figurew WITH wordr-figurew         (c23-c37);   !
 visual-flags WITH general-figurew        (c38-c97);   ! Cross-covariances
 general-wordm WITH addition-figurew     (c98-c147);   !
 addition-straight  WITH  wordr-figurew (c148-c171); !

MODEL CONSTRAINT:                                       ! Model constraint
 NEW(j1-j4 vv1-vv4 cc1-cc4 a1-a4 vt ct sa);             ! New parameters
 j1=4; j2=5; j3=4; j4=6;                                ! Number of items
 vv1=sum(v1-v4);  vv2=sum(v5-v9);                       ! Subscale variance
                                                          sums
 vv3=sum(v10-v13); vv4=sum(v14-v19);                    !
 cc1=sum(c1-c6);  cc2=sum(c7-c16);                      ! Subscale
                                                          covariance sums
 cc3=sum(c17-c22); cc4=sum(c23-c37);                    !
 DO (1,4) a#=(j#/(j#-1))*(1-vv#/(vv#+2*cc#));           ! Subscale alphas
 vt=sum(v1-v19);                                        ! Total variances
 ct=sum(c1-c171);                                       ! Total covariances
 sa=1-((vv1+2*cc1)*(1-a1)+(vv2+2*cc2)*(1-a2)            ! Stratified alpha
 +(vv3+2*cc3)*(1-a3)+(vv4+2*cc4)*(1-a4))/(vt+2*ct)); !
```

The resulting relevant output is:

```
                    Estimate       S.E.  Est./S.E.    P-Value
...
    A1                 0.653      0.038     16.980      0.000
    A2                 0.839      0.015     54.274      0.000
    A3                 0.754      0.030     25.245      0.000
    A4                 0.694      0.035     19.612      0.000
...
    SA                 0.861      0.016     53.281      0.000
```

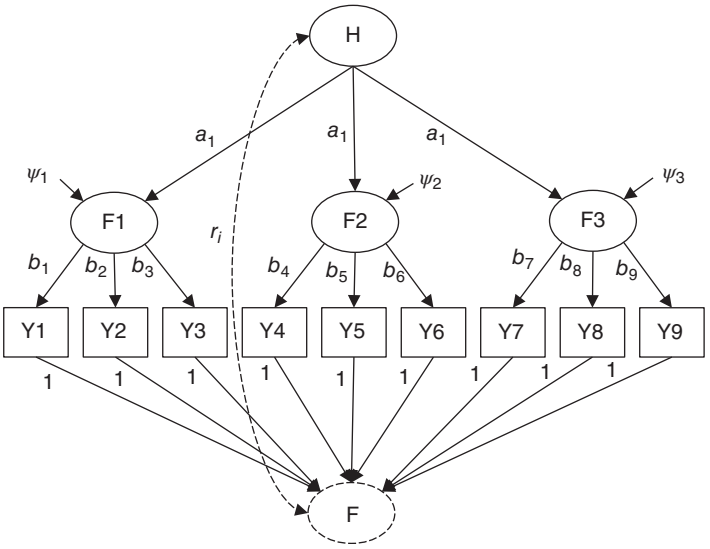


Figure 3.10 Reliability index for the correlated factors model confirmatory factor analysis.

The individual subscale alphas range from 0.653 to 0.839, with a composite stratified alpha value of 0.861. Like coefficient alpha for unidimensional data, stratified alpha for multidimensional data assumes tau-equivalence or equality of the loadings for each factor, as well as the absence of correlated errors.

6.2 Reliability Index with the Correlated Factors Model

This model includes a unit weighted composite of the observed scores (F0) and a latent composite of the three common factors (F4) which is weighted by the sum of the corresponding indicator factor loadings (see Raykov & Shrout, 2002). The reliability index (r_i) is the model implied correlation between the observed score composite and the latent variable composite (Figure 3.10). The squared reliability index is the reliability of the observed score composite in estimating the latent variable composite. The following Mplus syntax [RICFM19.inp] can be used to obtain the reliability index for this model and other models with a more general structure.

```
...
USEVARIABLES ARE visual-figurew;          ! Select 19 variables
DEFINE: STANDARDIZE visual-figurew;       ! Standardize observed
variables
MODEL:                                    ! Model
 spatial BY visual-flags*      (b1-b4);   ! Common factors
 verbal  BY general-wordm*     (b5-b9);   !
 speed   BY addition-straight* (b10-b13); !
 memory  BY wordr-figurew*     (b14-b19); !
 spatial-memory@1;                        ! Fix factor variances to 1
 F BY; G BY;                              ! Composites
```

```
 F-G@0; F WITH G@0;                          ! Fix to 0
 F ON visual-figurew@1;                      ! Observed variable composite
 G ON spatial-memory (a1-a4);                ! Latent variable composite
 MODEL CONSTRAINT:                           ! Model constraint
 a1=sum(b1-b4);                              ! Factor loading sums
 a2=sum(b5-b9);                              !
 a3=sum(b10-b13);                            !
 a4=sum(b14-b19);                            !
OUTPUT: STDYX; TECH4;                        ! Model implied correlations
```

We have shown the estimated reliability index in bold font in this output sample.

```
          ESTIMATED CORRELATION MATRIX FOR THE LATENT VARIABLES
              SPATIAL      VERBAL      SPEED       MEMORY        F
 SPATIAL        1.000
 VERBAL         0.594       1.000
 SPEED          0.583       0.467       1.000
 MEMORY         0.635       0.496       0.605       1.000
 F              0.793       0.780       0.752       0.796       1.000
 G              0.831       0.817       0.788       0.835       0.954
```

Squaring the reliability index (0.954) gives the estimated reliability of 0.910. Using the squared reliability index gives the point estimate of the reliability, but the output does not provide a direct estimate of the uncertainty associated with the reliability estimate, nor does the procedure shown here permit bootstrapping of the reliability index or the reliability estimate (Figure 3.11).

The following Mplus syntax [RICFM19_SUB.inp] can be used to obtain factor-specific reliability index values for the four subscales from the first 19 tests in the Holzinger–Swineford data.

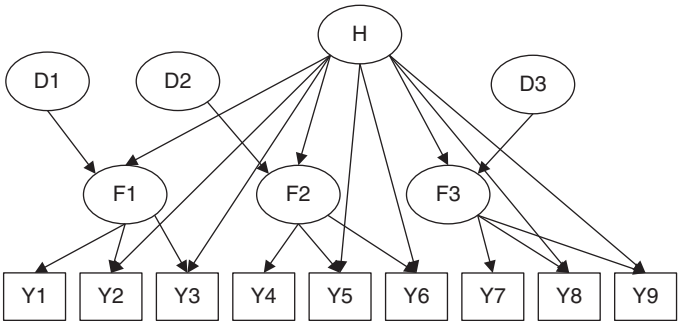


Figure 3.11 Factor-specific reliability indices for the correlated factors model.

```
...
USEVARIABLES ARE visual-figurew;            ! Select 19 variables
DEFINE: STANDARDIZE visual-figurew;         !
MODEL: spatial BY visual-flags* (b1-b4);    ! Common factors
 verbal  BY general-wordm*      (b5-b9);    !
 speed   BY addition-straight*  (b10-b13);  !
 memory  BY wordr-figurew*      (b14-b19);  !
 spatial-memory@1;                          ! Fix factor variances to 1
 T1 BY; T2 BY; T3 BY; T4 BY;                ! Subscales
 T1-T4@0;  T1-T4 WITH T1-T4@0;              ! Fix to zero
 T1 ON visual-flags@1;                      ! Subscale scores
 T2 ON general-wordm@1;                     !
 T3 ON addition-straight@1;                 !
 T4 ON wordr-figurew@1;                     !
OUTPUT: STDYX; TECH4;                       ! Latent variable
                                              correlations
```

The corresponding Mplus output provides the estimated reliability index correlations between the factors and the factor-specific composite scores.

```
          ESTIMATED CORRELATION MATRIX FOR THE LATENT VARIABLES
             SPATIAL       VERBAL        SPEED         MEMORY
...
 T1           0.832         0.494         0.485         0.528
 T2           0.563         0.949         0.443         0.471
 T3           0.519         0.415         0.890         0.539
 T4           0.546         0.427         0.520         0.860
```

The factor and subscale specific reliabilities for the correlated factors model can be obtained by squaring the correlations between the factors and the corresponding unit-weighted composite measures. The reliabilities are then 0.692, 0.901, 0.792, and 0.740 for the factors Spatial, Verbal, Speed, and Memory, respectively.

6.3 Reliability Index for the Higher Order Factor Model

This model includes a unit weighted composite of the observed scores (F) and a higher-order factor (H) using the following Mplus syntax [RIHO19.inp] (Figure 3.12).

```
...
USEVARIABLES ARE visual-figurew;          ! Select 19 variables
DEFINE: STANDARDIZE visual-figurew;      !  Standardize observed variables
MODEL: spatial BY visual-flags* (b1-b4);  ! First-order factors
 verbal  BY general-wordm*      (b5-b9);  !
 speed   BY addition-straight*  (b10-b13);!
 memory  BY wordr-figurew*      (b14-b19);!
 H BY spatial-memory* (a1-a4); H@1;       ! Second order factor
 spatial-memory (p1-p4);                  ! First-order residual variances
 F BY; F@0                                ! Composites
 F on visual-figurew@1;                   ! Observed variable composite
 MODEL CONSTRAINT: DO (1,4) p#=1-a#^2;    ! First-order residual variances
OUTPUT: STDYX; TECH4;                     ! Model implied correlations
```

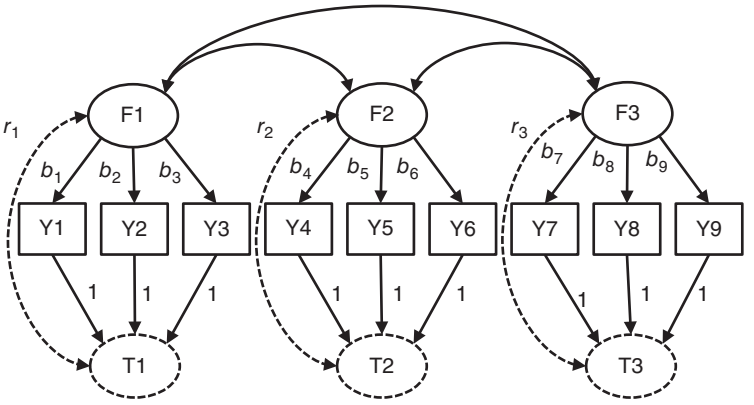


Figure 3.12 Reliability index for the higher order confirmatory factor analysis.

In this Mplus syntax, we select and standardize the first 19 variables from the Holzinger–Swineford data. The higher factor is identified and scaled by fixing its variance to 1. The four first-order factors are identified and scaled by fixing their residual variances to $1-a_k^2$ where a_k is the second order factor loading of the kth first order factor on the higher order factor H. The observed score F is the unit-weighted composite of all the items. The reliability index for the higher order model is the model implied correlation between the higher order factor and the observed total composite score and is shown in bold in the following Mplus output.

```
...
          ESTIMATED CORRELATION MATRIX FOR THE LATENT VARIABLES
               SPATIAL     VERBAL      SPEED       MEMORY      H
 SPATIAL         1.000
 VERBAL          0.560       1.000
 SPEED           0.610       0.485       1.000
 MEMORY          0.651       0.518       0.564       1.000
 H               0.839       0.668       0.727       0.776       1.000
 F               0.791       0.783       0.751       0.797       0.865
```

Squaring the reliability index of 0.865 gives the estimated reliability of 0.748.

6.4 Reliability Estimation Via the Reliability Index in the Bifactor CFA

The reliability index can be defined as the correlation between a unit-weighted composite score and an appropriate corresponding latent variable. In the case of the bifactor CFA, the correlation of a composite consisting of the sum of all of the items with the general factor from the bifactor CFA model, is an estimated reliability index whose squared value is the commonly used reliability coefficient for the overall component of multidimensional measures, namely Omega hierarchical (ω_H). Similarly, the correlations of the specific factors from the bifactor CFA with corresponding unit-weighted subscale scores provides reliability index estimates for each of the subscales. The squared reliability index values provide a reliability estimate, known as Omega hierarchical subscale (ω_{H_s}) for each subscale.

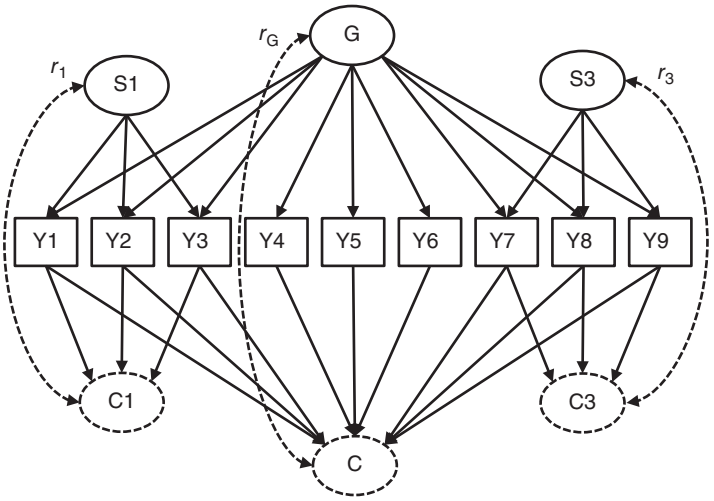


Figure 3.13 Reliability indices for the omit one specific factor (S-1) Bifactor Model.

In Figure 3.13, r_{GC} is the model implied correlation between the general factor, G, and the unit weighted composite of all the items, C. This is equal to the reliability index for the general factor and the overall composite, and its squared value r^2_{GC} is the reliability coefficient also known as Omega hierarchical ω_H. The model implied correlation r_1 is the model implied correlation between the first specific factor, S1, and the unit weighted composite of the first three items, C1. This is the reliability index for the first specific factor and the first subscale score and its squared value r^2_1 is the reliability coefficient also known as Omega hierarchical subscale ω_{Hs}. Similarly, r_3 and its squared value r^2_3 are the reliability index and coefficient for the last specific factor and subscale score.

In Mplus, this can be obtained from the following syntax [RIBI24.inp]:

```
...
! Bifactor CFA Reliability estimation: The reliability index
USEVARIABLES ARE visual-arithmet;    ! Select 24 variables
ANALYSIS: MODEL IS NOCOVARIANCES;    ! Fix factor covariances to 0
MODEL: G BY visual-arithmet;         ! General factor
 spatial BY visual-flags;            ! Specific factors
 verbal  BY general-wordm;           !
 speed   BY addition-straight;       !
 memory  BY wordr-figurew;           !
 F0 BY; F1 BY; F2 BY; F3 BY; F4 BY; ! Composite variables
 F0-F4@0;                            ! Fix variances to 0
 F0 ON visual-arithmet@1;            ! Unit-weighted composites
 F1 ON visual-flags@1;               !
 F2 ON general-wordm@1;              !
 F3 ON addition-straight@1;          !
 F4 ON wordr-figurew@1;              !
OUTPUT: STDYX; TECH4;                ! Standardized latent correlations
```

The output selection containing selected estimated correlations of the total composite (F0) and the subscale scores (F1–F4) with the general factor (G) and the specific latent variables (SPATIAL, VERBAL, SPEED, and MEMORY) is given below. In the upper left of this output snippet, the correlation between F0 and G is 0.875, the reliability index for the total score and the general factor. Squaring this value gives the estimated Omega hierarchical reliability coefficient ($\omega_H = 0.766$) for the whole scale. Similarly, the reliability index values for the corresponding subscale scores and specific factors can be found in the remaining diagonal entries, and give the reliability index values for each of the subscales. For example, the correlation between F1 and SPATIAL is 0.495, when squared giving ($\omega_{H_1} = 0.245$) as the Omega hierarchical subscale coefficient estimate for the first subscale.

```
      ESTIMATED CORRELATION MATRIX FOR THE LATENT VARIABLES
            G             SPATIAL       VERBAL        SPEED         MEMORY
...
 F0       0.875         0.057         0.121         0.337         0.098
 F1       0.699         0.495         0.000         0.000         0.000
 F2       0.698         0.000         0.630         0.000         0.000
 F3       0.625         0.000         0.000         0.630         0.000
 F4       0.629         0.000         0.000         0.000         0.564
```

Although this approach allows direct estimation of the Omega hierarchical and Omega hierarchical subscale reliability coefficients without computations on the model factor loadings, it requires the creation of "phantom" latent composite variables for the overall measure and each subscale, and it requires squaring the reliability index correlations to obtain the appropriate reliability coefficients.

6.5 Reliability Estimation for Higher-Order Factor Models Using Analytic Formulas

With a single higher order factor, an Omega-like reliability coefficient for the higher order factor with K first order factors each having J_k items can be defined as follows where λ_{jk} are the first order factor loadings of item j on factor k, γ_k are the second order factor loadings of first order factor k on the second order factor, ψ_k is the (residual) variance of first order factor k, and θ_{jj} are the indicator residual variances.

$$\omega_{HO} = \frac{\left(\sum_{k=1}^{K}\sum_{j=1}^{J_k}\lambda_{jk}\gamma_k\right)^2}{\left(\sum_{k=1}^{K}\sum_{j=1}^{J_k}\lambda_{jk}\gamma_k\right)^2 + \sum_{k=1}^{K}\psi_k\left(\sum_{j=1}^{J_k}\lambda_{jk}\right)^2 + \sum_{j=1}^{J}\theta_{jj}}$$

The Mplus syntax for this [OMEGA_HO19.inp] is:

```
...
USEVARIABLES ARE visual-figurew;           ! Select 19 variables
DEFINE: STANDARDIZE visual-figurew;        ! Standardize observed
                                             variables
MODEL: spatial BY visual-flags* (b1-b4);   ! First-order factors
 verbal  BY general-wordm*      (b5-b9);   !
 speed   BY addition-straight*  (b10-b13); !
 memory  BY wordr-figurew*      (b14-b19); !
 H BY spatial-memory* (a1-a4); H@1;        ! Second order factor
 spatial-memory (p1-p4);                   ! First-order residual
                                             variances
 visual-figurew (e1-e19);                  ! Residual variances
MODEL CONSTRAINT: DO (1,4) p#=1-a#^2;      ! Identify FO factors
NEW (x y z o);                             ! New parameters
 x=(a1*b1+a1*b2+a1*b3+a1*b4+               ! Second order
    a2*b5+a2*b6+a2*b7+a2*b8+a2*b9+         !
    a3*b10+a3*b11+a3*b12+a3*b13+           !
    a4*b14+a4*b15+a4*b16+a4*b17+           !
    a4*b18+a4*b19)^2;                      !
 y= p1*(b1+b2+b3+b4)^2+                    ! First order
    p2*(b5+b6+b7+b8+b9)^2+                 !
    p3*(b10+b11+b12+b13)^2+                !
    p4*(b14+b15+b16+b17+b18+b19)^2;        !
 z= sum(e1-e19);                           ! Residual variance sum
 o= x/(x+y+z);                             ! Higher order reliability
OUTPUT: STDYX; CINTERVAL;                  ! Confidence intervals
```

This Mplus syntax implements the analytic formula for higher order reliability and produces output containing the estimated reliability coefficient along with 90%, 95%, and 99% confidence intervals.

```
                    Estimate       S.E.  Est./S.E.    P-Value
New/Additional Parameters
...
    O                  0.749      0.034     21.728      0.000
```

```
         Lower .5% Lower 2.5%  Lower 5%  Estimate  Upper 5% Upper 2.5%  Upper .5%
New/Additional Parameters
...
    O       0.660     0.681     0.692     0.749     0.805     0.816     0.838
```

Here the reliability coefficient is estimated as a derived parameter rather than as a squared model-implied latent variable correlation. With Mplus, this approach allows the possibility of bootstrapping to get more appropriate estimates of the uncertainty and confidence intervals associated with the reliability coefficient.

7 Omega Reliability Estimation Via Reliability Formulas for Hierarchical Data

Another common way to obtain reliability coefficients for multidimensional measures is to use formulas for Omega hierarchical and other related reliability coefficients presented recently in Rodriguez, et al. (2016, *a*, *b*). Most commonly, these formulas are based on the factor loadings and residual variances from bifactor models. We have already met Omega hierarchical.

7.1 Omega Hierarchical

Omega Hierarchical is the proportion of the variability in the unit-weighted total scores that can be attributed to individual differences on the general factor.

$$\omega_H = \frac{\left(\sum \lambda_g\right)^2}{\left(\sum \lambda_g\right)^2 + \sum_{s=1}^{S}\left(\sum \lambda_s\right)^2 + \sum \theta}$$

In this equation, $(\sum \lambda_g)^2$ is the square of the sum of the loadings on the general factor, $(\sum \lambda_s)^2$ is the square of the sum of the specific factor loadings for a given subscale, $\sum_{s=1}^{S}(\sum \lambda_s)^2$ is the sum of these squared loadings across all of the specific factors and subscales, and $\sum \theta$ is the sum of the residual variances of all of the items. The numerator contains only the squared sum of the general factor loadings $(\sum \lambda_g)^2$. Given a bifactor model that fits the data adequately, the denominator of the equation is the total model-implied variance of the measure. Another reliability coefficient that is seen commonly with hierarchical or multidimensional data is known as Omega total.

7.2 Omega Total

Omega total, or sometimes simply Omega in the context of multidimensional or hierarchical measures, is a model-based estimate of the internal reliability of a unit-weighted multidimensional composite based on all of the items given both the general and the specific factors.

$$\omega_T = \frac{\left(\sum \lambda_g\right)^2 + \sum_{s=1}^{S}\left(\sum \lambda_s\right)^2}{\left(\sum \lambda_g\right)^2 + \sum_{s=1}^{S}\left(\sum \lambda_s\right)^2 + \sum \theta}$$

This equation differs from the one for Omega hierarchical by including the squared sum of the loadings for the specific factors $\sum_{s=1}^{S}(\sum \lambda_s)^2$ along with the squared sum of the general factor loadings in the numerator of the equation. Both the general and specific components are modeled as representing the contribution of true score variance. It can easily be seen that ω_T will exceed ω_H whenever there is systematic variance due to the specific factors in the model, that is, whenever $\sum_{s=1}^{S}(\sum \lambda_s)^2 > 0$. Omega hierarchical and Omega total can be useful reliability coefficients for the full set of items from multidimensional measures. Both coefficients use the same total variance estimate in the denominator of the equation. However, there is sometimes additional direct interest in each of the specific factors and the corresponding subscales. In this context, the specific factors have the variance they share with the rest of the items through the general factor controlled for or partialled out. Again, we have already met Omega hierarchical subscale which indicates how reliable a subscale is as a measure of a given specific factor.

7.3 Omega Hierarchical Subscale

Omega hierarchical subscale gives the specific factor variance relative to the general factor, specific factor, and residual variance of the items on the subscale.

$$\omega_{H_s} = \frac{(\sum \lambda_s)^2}{(\sum \lambda_{g_s})^2 + (\sum \lambda_s)^2 + \sum \theta_s}$$

In this equation, $(\sum \lambda_{g_s})^2$ is the squared sum of the general factor loadings for the items on a given subscale and $\sum \theta_s$ is the sum of the residual variances for the subscale items. The denominator of the equation is the model based estimate of the total variance of the subscale score for a given subscale.

7.4 Omega Subscale

Omega subscale is the proportion of the variability in the unit weighted subscale item scores that can be attributed to either the general factor or the specific factor.

$$\omega_{S_s} = \frac{(\sum \lambda_{g_s})^2 + (\sum \lambda_s)^2}{(\sum \lambda_{g_s})^2 + (\sum \lambda_s)^2 + \sum \theta_s}$$

This equation differs from the equation for Omega hierarchical subscale by including the squared sum of the general factor loadings for the items in the subscale along with the squared sum of subscale loadings in the numerator of the equation. The denominator of both equations is the estimated total variance for a given subscale. Both the general and specific components of the subscale are modeled as contributing and representing true score variance. It can easily be seen that ω_{S_s} will exceed ω_{H_s} whenever there is systematic variance due to the general factor for the items in the subscale, that is, whenever $(\sum \lambda_{g_s})^2 > 0$.

7.5 Calculating Omega Reliability Coefficients from Factor Loadings and Residual Variances

Mplus and other software that allows the creation of linear and nonlinear combinations of model parameters can be used to directly compute the Omega coefficients mentioned above. In doing so, we use the unstandardized model parameters. While the results from using standardized and unstandardized parameters are generally quite similar, it is generally more convenient to work with the unstandardized ones, as they permit more direct computation of parameter confidence intervals, the possible use of bootstrapping, etc. Calculating reliability coefficients based on unstandardized parameters can sometimes be more readily interpreted in terms of the metric of the original variables. If these variables have meaningful scales, either individually or relative to one another, it can be useful to create composite scores of these original variables rather than first converting them to z-scores in order to do the analysis. Still, it is common to use standardized parameters and we will illustrate later how to do this.

The following Mplus syntax can be used to calculate Omega hierarchical, Omega total, Omega hierarchical subscale, and Omega subscale. In our strategy, we make use of (many) intermediate quantities as building blocks for implementing the necessary standard formulas which are given in Rodriguez, et al. (2016 *a*, *b*). We use all 24 of the mental tests from the Holzinger–Swineford data, separately labeling the factor loadings for the general and specific factors in the bifactor CFA model

where there are some indicators that directly reflect the general factor. We also label the indicator residual variances for use in reliability formula construction. We then sum the relevant factor loadings separately for the different item subsets that constitute the designated subscales. Based on these and a few other quantities, we implement the necessary reliability formulas, again, using Mplus' MODEL CONSTRAINT: command in the following Mplus syntax [OMEGA_BI24.inp].

```
...
USEVARIABLES ARE visual-arithmet;    ! Include all 24 variables
ANALYSIS: MODEL IS NOCOVARIANCES;    ! Omit covariances
MODEL: G BY                          ! General factor
 visual-flags*            (g1-g4)    ! General factor loadings
 general-wordm*           (g5-g9)    !
 addition-straight*       (g10-g13)  !
 wordr-figurew*           (g14-g19)  !
 deduct-arithmet*         (g20-g24); !
 f1 BY visual-flags*      (h1-h4);   ! Specific factor loadings
 f2 BY general-wordm*     (h5-h9);   !
 f3 BY addition-straight* (h10-h13); !
 f4 BY wordr-figurew*     (h14-h19); !
 g@1 f1-f4@1;                        ! Fix factor variances to one
 visual-arithmet          (r1-r24);  ! Residual variances
OUTPUT: STDYX; CINTERVAL;            ! Standardized estimates

MODEL CONSTRAINT:               ! Model constraints
 NEW(fg1-fg5 fh1-fh4 fr1-fr5 g r t); ! New variables
 fg1=sum(g1-g4);                ! Sums of general factor loadings
 fg2=sum(g5-g9);                !
 fg3=sum(g10-g13);              !
 fg4=sum(g14-g19);              !
 fg5=sum(g20-g24);              !
 fh1=sum(h1-h4);                ! Sums of specific factor loadings
 fh2=sum(h5-h9);                !
 fh3=sum(h10-h13);              !
 fh4=sum(h14-h19);              !
 fr1=sum(r1-r4);                ! Sums of residual variances
 fr2=sum(r5-r9);                !
 fr3=sum(r10-r13);              !
 fr4=sum(r14-r19);              !
 fr5=sum(r20-r24);              !
 g=sum(g1-g24);                 ! Sum of all general factor loadings
 r=sum(r1-r24);                 ! Sum of all residual variances
 t=g^2+fh1^2+fh2^2+fh3^2+fh4^2+r; ! Total variance estimate

NEW(omegat omegah os1-os4 oh1-oh4);            ! Omega coefficients
 omegat=(g^2+fh1^2+fh2^2+fh3^2+fh4^2)/t;       ! Omega total
 omegah=g^2/t;                                 ! Omega hierarchical
 DO (1,4) os#=(fg#^2+fh#^2)/(fg#^2+fh#^2+fr#); ! Omega subscale
 DO (1,4) oh#=fh#^2/(fg#^2+fh#^2+fr#);         ! Omega hierarchical
                                                 subscale
```

The results are included in the Mplus output as New/Additional Parameters.

```
New/Additional Parameters
                    Estimate       S.E.  Est./S.E.    P-Value
...
    OMEGAT             0.906      0.012     75.056      0.000
    OMEGAH             0.766      0.034     22.727      0.000
    OS1                0.734      0.049     14.837      0.000
    OS2                0.885      0.017     53.094      0.000
    OS3                0.788      0.030     26.466      0.000
    OS4                0.713      0.042     17.147      0.000
    OH1                0.245      0.071      3.454      0.001
    OH2                0.397      0.065      6.068      0.000
    OH3                0.397      0.068      5.847      0.000
    OH4                0.318      0.072      4.406      0.000
```

As expected, Omega total was quite high at 0.906, with an Omega hierarchical of 0.766. The Omega subscale values were also fairly high, ranging from 0.713 to 0.885, with Omega hierarchical subscale values ranging from 0.245 to 0.397. Although the significance tests are generally not meaningful because they test simply whether the coefficient differs from zero, the standard errors could be used to construct tests relative to certain cutoff values of interest. It's probably simpler in this case to just request confidence intervals. The results in the output below give the point estimate of the coefficient, along with 90%, 95%, and 99% confidence intervals. However, note that these are symmetric normal theory delta method confidence intervals that can have poor coverage because the estimated reliability coefficients seldom have normal distributions.

```
CONFIDENCE INTERVALS OF MODEL RESULTS
            Lower .5% Lower 2.5%  Lower 5%  Estimate  Upper 5% Upper 2.5% Upper .5%
...
    OMEGAT      0.875      0.883     0.887     0.906     0.926     0.930     0.938
    OMEGAH      0.679      0.700     0.710     0.766     0.821     0.832     0.852
    OS1         0.606      0.637     0.652     0.734     0.815     0.831     0.861
    OS2         0.842      0.852     0.858     0.885     0.912     0.918     0.928
    OS3         0.711      0.729     0.739     0.788     0.837     0.846     0.864
    OS4         0.606      0.632     0.645     0.713     0.782     0.795     0.820
    OH1         0.062      0.106     0.128     0.245     0.362     0.384     0.428
    OH2         0.229      0.269     0.290     0.397     0.505     0.526     0.566
    OH3         0.222      0.264     0.286     0.397     0.509     0.530     0.572
    OH4         0.132      0.177     0.199     0.318     0.437     0.460     0.504
```

In Mplus, it is relatively easy to request either bias corrected or ordinary percentile bootstrap confidence intervals. This is done by making some minimal modifications to the ANALYSIS: and OUTPUT: commands. For example, the following changes request 1000 bootstrap samples with bias-corrected bootstrap confidence intervals [OMEGA_BI24_BOOT.inp].

```
...
ANALYSIS: MODEL IS NOCOVARIANCES;    ! Omit covariances
 BOOTSTRAP=1000;                     ! 1000 bootstrap samples
...
OUTPUT: STDYX;                       ! Standardized estimates
 CINTERVAL(BCBOOTSTRAP) ;            ! Bootstrap confidence intervals
```

CONFIDENCE INTERVALS OF MODEL RESULTS ***(BCBOOTSTRAP)***

	Lower .5%	Lower 2.5%	Lower 5%	Estimate	Upper 5%	Upper 2.5%	Upper .5%
...							
OMEGAT	0.853	0.868	0.876	0.906	0.928	0.931	0.938
OMEGAH	0.638	0.669	0.684	0.766	0.829	0.840	0.860
OS1	0.519	0.576	0.609	0.734	0.846	0.949	1.337
OS2	0.796	0.839	0.844	0.885	0.909	0.915	0.924
OS3	0.667	0.703	0.721	0.788	0.831	0.838	0.851
OS4	0.538	0.599	0.613	0.713	0.782	0.795	0.942
OH1	0.002	0.062	0.098	0.245	0.354	0.392	0.601
OH2	0.173	0.224	0.256	0.397	0.573	0.599	0.673
OH3	0.189	0.243	0.266	0.397	0.526	0.563	0.609
OH4	0.042	0.117	0.158	0.318	0.446	0.474	0.630

The bootstrapping increases the execution time substantially, especially with our relatively inefficient method relying on so many intermediate parameter combination estimates. Using 1000 bootstrap samples with bias correction took over eight minutes to run, compared with only about a second or so without the bootstrapping.

8 Expected Common Variance

8.1 Expected Common Variance for the General Factor

Another reliability-related estimate for the bifactor model is the expected common variance (ECV). ECV is the proportion of reliable variance attributable to the general factor.

$$ECV = \frac{\sum \lambda_g^2}{\sum \lambda_g^2 + \sum_{s=1}^{S} \left(\sum \lambda_s^2 \right)}$$

In this equation, $\sum \lambda_g^2$ is the sum of the squared general factor loadings, and $\sum \lambda_s^2$ is the sum of the squared factor loadings for a particular subscale.

8.2 Expected Common Variance for Subscales

This is the proportion of reliable variance for the items on a particular subscale, that can be attributed to the specific factor.

$$ECV_s = \frac{\sum \lambda_s^2}{\sum \lambda_{g_s}^2 + \sum \lambda_s^2}$$

In this equation, $\sum \lambda_{g_s}^2$ is the sum of the squared general factor loadings for the items on a particular subscale.

8.3 Item Expected Common Variance

This is the proportion of reliable variance in an item that is attributable to the general factor.

$$ECV_I = \frac{\lambda_{g_j}^2}{\lambda_{g_j}^2 + \lambda_{s_j}^2}$$

For items with loadings on both the general factor and a specific factor, $\lambda_{g_j}^2$ is the squared general factor loading for item j, and $\lambda_{s_j}^2$ is the squared specific factor loading for this item. For items reflecting only the general factor, the value of I-ECV becomes 1.0, and I-ECV is undefined for items loading only on the specific factor. The ECV coefficients for the general and specific factors, and for items, can be obtained with the following Mplus syntax [ECV_BI24.inp].

```
...
USEVARIABLES ARE visual-arithmet;          ! Include all 24 variables
ANALYSIS: MODEL IS NOCOVARIANCES;          ! Omit covariances
MODEL: G BY                                ! General factor
 visual-flags*            (g1-g4)          ! General factor loadings
 general-wordm*           (g5-g9)          !
 addition-straight*       (g10-g13)        !
 wordr-figurew*           (g14-g19)        !
 deduct-arithmet*         (g20-g24);       !
 f1 BY visual-flags*      (h1-h4);         ! Specific factor loadings
 f2 BY general-wordm*     (h5-h9);         !
 f3 BY addition-straight* (h10-h13);       !
 f4 BY wordr-figurew*     (h14-h19);       !
 g@1 f1-f4@1;                              ! Fix factor variances to one
 visual-arithmet          (r1-r24);        ! Residual variances

MODEL CONSTRAINT: NEW(p1-p5 q1-q4 v);      ! Intermediate quantities
 p1=g1^2+g2^2+g3^2+g4^2;                   ! General factor variance
 p2=g5^2+g6^2+g7^2+g8^2+g9^2;              !
 p3=g10^2+g11^2+g12^2+g13^2;               !
 p4=g14^2+g15^2+g16^2+g17^2+g18^2+g19^2;   !
 p5=g20^2+g21^2+g22^2+g23^2+g24^2;         !
 q1=h1^2+h2^2+h3^2+h4^2;                   ! Specific factor variance
 q2=h5^2+h6^2+h7^2+h8^2+h9^2;              !
 q3=h10^2+h11^2+h12^2+h13^2;               !
 q4=h14^2+h15^2+h16^2+h17^2+h18^2+h19^2;   !
 v=sum(p1-p5 q1-q4);                       ! Total explained variance
NEW(ecvg ecv1-ecv4 iecv1-iecv19);          ! New parameters
 ecvg=(p1+p2+p3+p4+p5)/v;                  ! General factor ECV
 DO (1,4) ecv#=q#/(p#+q#);                 ! Subscale ECV
 DO (1,19) iecv#=g#^2/(g#^2+h#^2);         ! Item ECV
OUTPUT: STDYX;                             ! Standardized estimates
```

The related output for the Holzinger–Swineford data is

```
                    Estimate       S.E.  Est./S.E.    P-Value
...
    ECVG               0.592      0.063      9.396      0.000
    ECV1               0.374      0.121      3.092      0.002
    ECV2               0.453      0.079      5.702      0.000
    ECV3               0.479      0.087      5.526      0.000
```

ECV4	0.473	0.112	4.229	0.000
IECV1	0.870	0.126	6.922	0.000
IECV2	0.584	0.219	2.669	0.008
IECV3	0.759	0.212	3.583	0.000
IECV4	0.508	0.181	2.799	0.005
IECV5	0.557	0.101	5.534	0.000
IECV6	0.484	0.098	4.926	0.000
IECV7	0.461	0.097	4.749	0.000
IECV8	0.758	0.099	7.626	0.000
IECV9	0.500	0.095	5.248	0.000
IECV10	0.303	0.116	2.606	0.009
IECV11	0.622	0.140	4.434	0.000
IECV12	0.333	0.110	3.031	0.002
IECV13	0.676	0.117	5.760	0.000
IECV14	0.471	0.196	2.400	0.016
IECV15	0.403	0.189	2.138	0.033
IECV16	0.756	0.149	5.067	0.000
IECV17	0.352	0.148	2.380	0.017
IECV18	0.699	0.157	4.440	0.000
IECV19	0.818	0.168	4.859	0.000

Although we have presented the Omega coefficients and the expected common variance measures separately, all of these estimates could be obtained in a single Mplus run. As with the Omega coefficients, confidence intervals and bootstrapping can be used to assess the uncertainty associated with the EVC measures. We have concentrated on obtaining and evaluating all of these reliability-related measures in an overall latent variable statistical modeling framework. We also note that there are a number of other measures discussed by Rodriguez, et al. (2016 *a*, *b*) that have been applied and found useful in assessing reliability of multidimensional measures.

9 Other Reliability Related Information

9.1 Construct Replicability Index

We have already encountered the construct replicability index in computational guide 1, where it was presented as the standardized coefficient of maximal reliability for unidimensional measures. In this equation, ℓ_j^2 are the standardized factor loadings.

$$H = 1/\left[1 + 1/\left(\sum_{j=1}^{J} \frac{\ell_j^2}{1-\ell_j^2}\right)\right]$$

9.2 Average Relative Parameter Bias

When assessing whether a traditional unidimensional model can adequately represent the general factor from a bifactor model for multidimensional data, it can be useful to compare these two sets of factor loadings. One measure of this discrepancy is the relative parameter bias. When there is

substantial evidence of multidimensionality, the general factor from the bifactor model is generally believed to represent the "truer" model. In this case the relative parameter bias for item j can be defined as the difference between the item's loading from the unidimensional single-factor model λ_{j_U} and its loading on the general factor from the bifactor model $\lambda_{j_{GB}}$, divided by the general factor loading.

$$RPB_j = \frac{\lambda_{j_U} - \lambda_{j_{GB}}}{\lambda_{j_{GB}}}$$

The average absolute parameter bias is the average of the absolute values of the relative parameter biases for all J items.

$$ARPB = \frac{\sum \left| RPB_j \right|}{J}$$

9.3 Factor Determinacy

Factor determinacy is the correlation between estimated factor scores and the corresponding latent variables.

In the following Mplus input syntax, all 24 variables are standardized by transforming them to variables with means of zero and standard deviations of 1. We estimate an orthogonal bifactor CFA by omitting covariances among the general and specific factors. The MODEL CONSTRAINT: command is used to estimate construct replicability for the general and specific factors and the average relative parameter bias (ARPB). To obtain ARPB, the standardized factor loadings from a preliminary single-factor model are brought in as new parameters. Because the MODEL CONSTRAINT: command does not provide a mathematical absolute value function, the same result was achieved by taking the square root of a squared quantity.

More generally, we note that Mplus' MODEL CONSTRAINT: command has somewhat primitive programming capabilities in comparison with other software for doing similar things. Without the availability of temporary variables, there are two primary choices: either build rather long and complicated expressions or proliferate intermediate parameter values. In the interest of transparently and concisely implementing and communicating formulas and equations for reliability coefficients, etc., we have chosen the latter approach. However, because of the comprehensive way Mplus treats all parameters, this can be relatively inefficient from a computational standpoint.

Finally, the FSDETERMINACY keyword is added to the OUTPUT: command to obtain the factor score determinacies for the general and specific factors [CR_BI24.inp].

```
USEVARIABLES ARE visual-arithmet;     ! Select all 24 variables
DEFINE: STANDARDIZE visual-arithmet; ! Standardize variables
ANALYSIS: MODEL IS NOCOVARIANCES;     ! Omit covariances
MODEL:                                ! Model
 G  BY visual-arithmet*   (g1-g24);   ! General factor loading
 F1 BY visual-flags*      (h1-h4);    ! Specific factor loadings
 F2 BY general-wordm*     (h5-h9);    !
 F3 BY addition-straight* (h10-h13); !
 F4 BY wordr-figurew*     (h14-h19); !
 G@1 F1-F4@1;                         ! Fix factor variances to 1
 MODEL CONSTRAINT:                    ! Model constraint
```

```
 NEW(p1-p24 q1-q19 pg qs1-qs4);       ! New parameters
 DO (1,24) p#=g#^2/(1-g#^2);          ! General factor item components
 DO (1,19) q#=h#^2/(1-h#^2);          ! Specific factor item components
 pg=sum(p1-p24);                      ! Sum of general components
 qs1=sum(q1-q4);                      ! Sums of specific components
 qs2=sum(q5-q9);                      !
 qs3=sum(q10-q13);                    !
 qs4=sum(q14-q19);                    !
 NEW(HG HS1-HS4);                     ! Construct replicability
 HG=1/(1+1/pg);                       ! General factor replicability
 DO (1,4) HS#=1/(1+1/qs#);            ! Specific factor replicability
 NEW(b1-b24 r1-r24);
 b1=.564; b2=.371; b3=.432; b4=.537; b5=.720; b6=.707;
 b7=.709; b8=.693; b9=.723; b10=.434; b11=.528; b12=.435;
 b13=.575; b14=.407; b15=.366; b16=.476; b17=.443; b18=.482;
 b19=.424; b20=.620; b21=.580; b22=.620; b23=.686; b24=.628;
 DO (1,24) r#=sqrt(((b#-g#)/g#)^2);  ! Item relative paramer bias
 NEW(ARPB); ARPB=mean(r1-r24);        ! Average relative parameter bias
OUTPUT: STDYX;                        ! Standardized etimate
 FSDETERMINACY;                       ! Factor score determinacy
```

The relevant output for the Holzinger–Swineford data is shown below

```
                    Estimate       S.E.  Est./S.E.    P-Value
...
    HG                 0.915      0.017     55.323      0.000
    HS1                0.401      0.119      3.381      0.001
    HS2                0.693      0.053     13.069      0.000
    HS3                0.622      0.068      9.205      0.000
    HS4                0.529      0.073      7.216      0.000
...
    ARPB               0.082      0.039      2.116      0.034
...
           FACTOR DETERMINACIES

           G           0.939
           F1          0.686
           F2          0.860
           F3          0.814
           F4          0.740
```

We have shown how to use Mplus to obtain various kinds of reliability-related information from multidimensional measures via functions of the parameters of the bifactor model. We have argued that this approach allows consideration of the uncertainty or confidence intervals associated with these entities, and to conduct inference concerning them. In many cases, however, the primary interest is simply in the point estimate of, say, the Omega hierarchical reliability coefficient, and it can be simple and convenient to use a spreadsheet or other computing application for this

purpose. The expressions in Rodriguez, et al. (2016 *a*, *b*) can be implemented directly by users if desired.

Alternatively, an Excel spreadsheet (Dueber, 2017) and R package (Dueber, 2020) have been made available by David Dueber for calculating the various mathematical and statistical indices for multidimensional data based on the results of the bifactor model. The R package can import lavaan or mirt objects from R, or can use Mplus CFA or ESEM output files. The R package can also provide certain of the indices for correlated factor, higher order factor, and multi-trait multi-method models, as well as bifactor models. If uncertainty or inference are not of interest, we recommend using Duebner's tools which can dramatically simplify the process.

10 Special Situations in Reliability Assessment

10.1 Negative Factor Loadings on the Specific Factors

For computing reliability coefficients such as Omega hierarchical, most bifactor CFA models exhibit the fortuitous property that all of the factor loadings are strictly positive. In the traditional one-factor model, or in the multidimensional correlated factors model, a negative factor loading is usually treated as evidence that an item or indicator should be reverse scored. A somewhat more ambiguous situation is encountered when all of the items on a particular factor have negative loadings. Because the sign of the latent variable can be indeterminate in some models, it may be that the scale of factor, and not the indicators, should be reversed to match theoretical expectations. This anomaly is encountered with some frequency in factor analysis. It can sometimes be resolved by using starting values or actually constraining the loadings to be positive.

In bifactor models, an even more puzzling situation can arise when all of the loadings on the general factor are positive, as expected, but there are approximately equal numbers of positive and negative loadings on one of the specific or group factors. Sometimes this can be due to under-extraction of the specific factors. It can also be due to trying to fit a bifactor model when the true model is a correlated factors model. At other times, however, the bifactor model is the correct model and the designated specific factor may be genuinely bipolar, with both true positive and true negative loadings. There may also be a plausible theoretical or methodological rationale to accompany the observed empirical findings. In this case, reversal of the items with negative specific factor loadings is incompatible with the strictly positive general factor loadings.

Counterbalancing of the positive and negative factor loadings on the specific factor can will reduce the specific factor's variance in the calculation of the usual Omega coefficients. Omega hierarchical will be over-estimated because the denominator of the coefficient will be underestimated, and Omega hierarchical subscale will be underestimated because of the shrinkage in the sum of the specific factor loadings. Morin et al. (2020) has proposed taking the absolute values of the factor loadings instead of their original signed values when creating the sums of the loadings for estimating the variance components in Omega coefficients. For example, Omega hierarchical is then expressed using the following equation.

$$\omega_{HA} = \frac{\left(\sum|\lambda_g|\right)^2}{\left(\sum|\lambda_g|\right)^2 + \sum_{s=1}^{S}\left(\sum|\lambda_s|\right)^2 + \sum\theta}$$

If all of the loadings are positive, the results of using this expression will be identical to those from using the original equation for Omega hierarchical. If some of the loadings are negative, however,

the results will differ. A similar situation occurs if some of the specific factor loadings are negative when calculating Omega hierarchical subscale.

$$\omega_{HA_s} = \frac{(\sum|\lambda_s|)^2}{(\sum|\lambda_{g_s}|)^2 + (\sum|\lambda_s|)^2 + \sum\theta_s}$$

10.2 Correlated Indicator Residuals

The violation of local independence given the latent variable structure is a potential concern with multidimensional data just as it is with unidimensional data. For unidimensional data, the Rho coefficient for unidimensional data generalizes the unidimensional Omega coefficient to the case with correlated residuals. In a similar way, we might generalize Omega hierarchical to Rho hierarchical as the multidimensional counterparts to their unidimensional coefficients. As with Omega hierarchical this is done using the factor loadings and residual variances and covariances of the bifactor model.

$$\rho_H = \frac{(\sum\lambda_g)^2}{(\sum\lambda_g)^2 + \sum_{s=1}^{S}(\sum\lambda_s)^2 + \sum\theta_{jj} + 2\sum\theta_{jj'}}$$

In this equation θ_{jj} is the residual variance for item j, and $\theta_{jj'}$ is a modeled non-zero residual covariance between two items, item j and item j'.

Similarly, Omega hierarchical subscale can be generalized to Rho hierarchical subscale by incorporating the within-subscale residual covariances. As with Rho hierarchical, in the equation for Rho hierarchical subscale, the $\theta_{(s)jj}$ give the within-subscale residual variances for subscale s and $\theta_{(s)jj'}$ give the modeled non-zero within-subscale residual covariances between item j and item j' on subscale s (Table 3.15).

$$\rho_{H_s} = \frac{(\sum\lambda_s)^2}{(\sum\lambda_{g_s})^2 + (\sum\lambda_s)^2 + \sum\theta_{(s)jj} + 2\sum\theta_{(s)jj'}}$$

Table 3.15 Schmid-Leiman transformed loadings for a higher order CFA with direct indicators of the higher order factor using the constrained bifactor and higher order models.

	Constrained bifactor		Higher order	
	Direct effects		Indirect effects	
	GENERAL	SPECIFIC	GENERAL	SPECIFIC
SPATIAL				
VISUAL	0.558	0.413	0.558	0.413
CUBES	0.404	0.299	0.404	0.299
PAPER	0.404	0.299	0.404	0.299
FLAGS	0.553	0.409	0.553	0.409

(*Continued*)

Table 3.15 (Continued)

	Constrained bifactor		Higher order	
	Direct effects		Indirect effects	
	GENERAL	SPECIFIC	GENERAL	SPECIFIC
VERBAL				
GENERAL	0.591	0.548	0.591	0.548
PARAGRAP	0.598	0.555	0.598	0.555
SENTENCE	0.613	0.569	0.613	0.569
WORDC	0.512	0.476	0.512	0.476
WORDM	0.620	0.575	0.620	0.575
SPEED				
ADDITION	0.460	0.489	0.460	0.489
CODE	0.464	0.494	0.464	0.494
COUNTING	0.493	0.524	0.493	0.524
STRAIGHT	0.503	0.535	0.503	0.535
MEMORY				
WORDR	0.400	0.339	0.400	0.339
NUMBERR	0.393	0.334	0.393	0.334
FIGURER	0.462	0.392	0.462	0.392
OBJECT	0.477	0.405	0.477	0.405
NUMBERF	0.480	0.407	0.480	0.407
FIGUREW	0.382	0.324	0.382	0.324
ABILITY				
DEDUCT	0.653		0.653	
NUMERIC	0.639		0.639	
PROBLEMR	0.653		0.653	
SERIES	0.732		0.732	
ARITHMET	0.636		0.636	

References

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: appreciating benefits and limitations. *Biological Psychiatry*, *88*(1), 18–27.

Burns, G. L., Geiser, C., Servera, M., Becker, S. P., & Beauchaine, T. P. (2020). Promises and pitfalls of latent variable approaches to understanding psychopathology: reply to Burke and Johnston, Eid, Junghänel and colleagues, and Willoughby. *Journal of Abnormal Child Psychology*, *48*(7), 917–922.

Cho, E. (2016). Making reliability reliable: a Systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651–682.

Dueber, D. M. (2017). Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models. https://dx.doi.org/10.13023/edp.tool.01. [Also available at https://education.uky.edu/edp/apslab/resources/]

Dueber, D. M. (2020). BifactorIndicesCalculator. R package version 0.2.1.

Eid, M. (2020). Multi-Faceted constructs in abnormal psychology: implications of the Bifactor S-1 model for individual clinical assessment. *Journal of Abnormal Child Psychology*, *48*(7), 895–900.

Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: the stability of a bi-factor solution. Supplementary Educational Monographs, 48, xi + 91.

Markon, K. E. (2019). Bifactor and hierarchical models: specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*, 51–69.

Morin, A. J. S., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and Bifactor-ESEM. In G. Tenenbaum, R. C. Eklund, & N. Boiangin (Eds.), Handbook of sport psychology: Exercise, methodologies, & special topics (4th ed., pp. 1044–1073). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119568124.ch51

Rae, G. (2007). A note on using stratified alpha to estimate the composite reliability of a test composed of interrelated nonhomogeneous items. *Psychological Methods*, *12*(2), 177–184.

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. Structural Equation Modeling, *9*(2), 195–212. https://doi.org/10.1207/S15328007SEM0902_3

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Evaluating bifactor models: calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237.

Online Datasets and Services

* Holzinger & Swineford dataset description and documentation is found at:
https://rdrr.io/cran/MBESS/man/HS.html

* Immediate direct download of the dataset as a csv file can be found found on the Wiley Companion Datasite for this book.

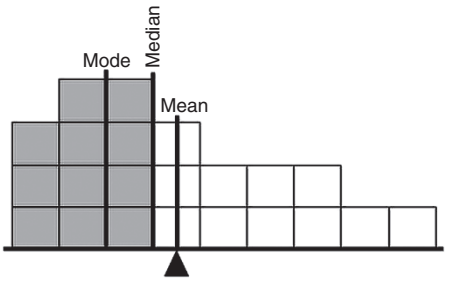
Appendix 1

Central Tendency: Mean, Median, and Mode

APPENDIX MENU

A1.1 Concepts and Formulas for Appendix 1, 487
A1.1.1 Concepts and Symbols, 487
A1.1.2 Formulas, 488
A1.2 Study Questions for Appendix 1, 488

In an elementary statistics book, the chapter on central tendency usually includes three statistics: the mean, the median, and the mode. It is instructive to compare these three to familiar physical quantities, to make each of them and their relative properties more memorable and understandable. Suppose there are 22 cubic wooden blocks that each measures precisely 1 inch to each side (height, width, and length). These are all stacked on a strong ruler of negligible weight, organized into a skewed distribution as shown in the first figure below. This distribution is clearly positively skewed. That is, its longest tail points to the right in a positive direction. Directly below the nine stacks of blocks which constitute this distribution is a row of values, X_i, that gives the scale value for each stack of blocks. Directly below this is a row of frequency values, f_i, that indicates how many blocks are in each of the stacks. The mode is defined as the data value that has the highest frequency, that is, the tallest stack of blocks. In this distribution, the scale values of 2 and 3 are tied with each having a frequency of 4, which is the largest of any of the 9 frequencies. Averaging the two scale values yields a mode of 2.5. Thus, the mode of this distribution is 2.5, the scale value corresponding to *the highest stack of blocks*.



Nine scale values (X_i):	1	2	3	4	5	6	7	8	9
Frequencies (f_i):	3	4	4	3	2	2	2	1	1
Products of scale values & frequencies ($X_i f_i$):	3	8	12	12	10	12	14	8	9

N = Sum of all 9 frequencies = $\Sigma f = 22$

Sum of all 22 scores = $\Sigma X_i = \Sigma X_i f_i = 88$

$$mean = \bar{X} = \frac{\Sigma X_i}{N} = \frac{88}{22} = 4.0$$

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

The median is the point on the scale that physically has the same number of blocks (scores) above it as below it. In this simple distribution, the median can be obtained by mere inspection, rather than calculation. It can be seen to be 3.5, as shown in the figure. There are eleven white-shaded blocks with scale values from 4 to 9 (above the median of 3.5), and 11 grey-shaded blocks with scale values from 1 to 3 (below the median of 3.5). Since each block is one cubic inch in size, the face of each is one square inch. The area of the face of this distribution is therefore 22 $in.^2$, 11 $in.^2$ below the median, and 11 $in.^2$ above the median. The median, in physical terms, is *the point that bisects the area* of the distribution. Notice that the median would not change with shifts in the shape of the distribution, as long as the 11 white blocks are all still below the median value of 3.5 and the 11 grey blocks are all above that median value. The median has a stability that resists reflecting local changes.

The mean is physically *the balance point of the distribution*. That is, with a distribution consisting of wooden blocks as described here, placed on a base with zero weight but strong enough to hold the blocks in place, the distribution would be perfectly balanced on a wedge placed at exactly the location of the arithmetic mean. The arithmetic mean, or the average as we commonly call it, is just the sum of all the values in the distribution divided by the number of observations in the distribution.

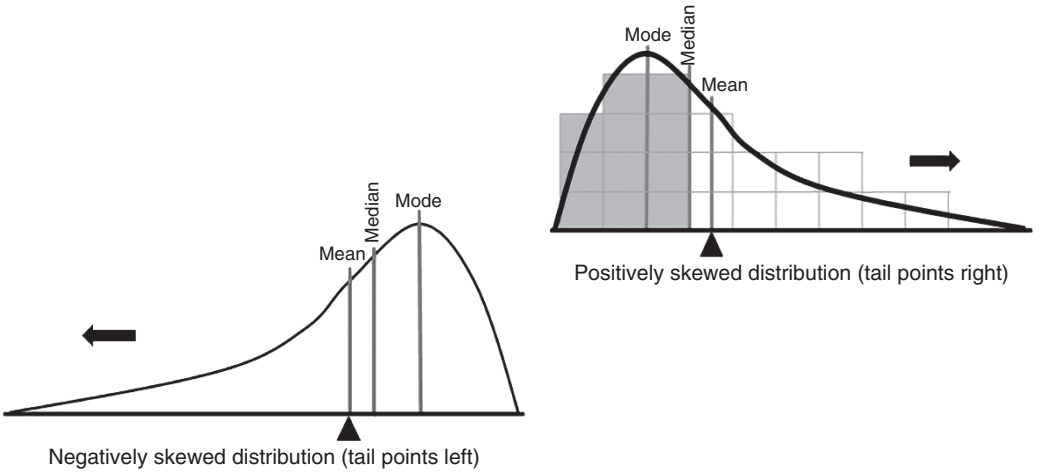
Shown immediately below the distribution is a simple process for efficiently calculating a mean from this graph of the distribution. After placing the scale values, X_i, in the first row directly below the graph, and then the frequency counts, f_i, for each scale value in a second row below that, a third row is created which contains the product of each scale value multiplied by the number of scores, f_i, that have that scale value. The sum of these 9 products yields the numerical sum of the scale values of all 22 scores, which turns out to be 88:

$$\sum X_i f_i = 3 + 8 + 12 + 12 + 10 + 12 + 14 + 8 + 9 = 88$$

Similarly, summing the frequency values, f_i, in the second row yields the total number of scores, which is $N = 22$. The mean is calculated by dividing the sum of the values of the scores by the number of observations, N, which gives us a mean of exactly 4.0.

$$mean = \bar{X} = \frac{\sum X_i}{N} = \frac{88}{22} = 4.0$$

Notice that the mean (4.0) is greater than the mode (2.5) along this horizontal scale, and the median (3.5) is in between these two values. This is usually the case (mean > median > mode) with a positively skewed distribution like this one, i.e. one with its tail pointing to the right, in a positive direction. The reverse will be true (mode > median > mean) for a negatively skewed distribution. To summarize, for a positively skewed distribution, the mean will usually be greater than the median, which will be greater than the mode; and for a negative distribution, the mode will usually be greater than the median, which will be greater than the mean.



©2020 John Wiley and Sons

Blahuš, P., Olsen, J. A., Bunderson, C. V., and Brown, B. L. *Psychometrics, Test Theory, and the Latent Factors Model*. Chapter 1 Appendix 1.1.

A1.1 Concepts and Formulas for Appendix 1

A1.1.1 Concepts and Symbols

Central tendency—The location of the "middle" of a distribution (mean, median, or mode) along a numerical scale.

Mean—The average of a set of values. Also, the expected value of a set of values. See formulas below.

Median—The point in the middle of a distribution of scores that has the same number of scores below it as above it. The percentile formula below can be adapted by setting P equal to 0.5 to calculate an "interpolated median."

Mode—The most frequently occurring score.

Skew or "skewness"—The property of a nonsymmetric distribution, indicating that it has a tail either pointing to the left (negative skew) or to the right (positive skew).

Symmetric distribution—A distribution that has no skew.

X_i—A variable. The "i" subscript indicates that the variable X can take multiple values. Y_i or W_i or another letter with subscripts could also be used to represent a variable. See "Summation sign with elaboration" below to see how the subscript functions.

Summation sign, Σ—This Greek letter sigma (capital sigma, to be exact) is the symbol for the mathematical operation of summation. Suppose you have six scores: 10, 8, 7, 4, 3, and 2. The summation sign indicates that you should add them up: $\Sigma X_i = 10 + 8 + 7 + 4 + 3 + 2 = 34$.

Summation sign with "elaboration," that is, with *detailed instructions:* $\sum_{i=1}^{N} X_i$—This summation sign is more explicit. It can specify with precision how you should summate and select from

the scores. Suppose, for example, the intent is simply to sum four of the six scores given in the preceding paragraph, from the second score to the fifth score. This would be the notation and the execution.

$$\sum_{i=2}^{5} X_i = X_2 + X_3 + X_4 + X_5 = 8 + 7 + 4 + 3 = 22$$

Scale value—This is a particular location along a numerical scale.

Score—This is the scale value corresponding to a particular person or object. Several or many persons or objects could each have a score corresponding to the same scale value.

A1.1.2 Formulas

Formulas for the mean:

$$\textit{The mean} = \bar{X} = \frac{\sum X_i}{N} \qquad \text{Second formula for the mean} = \bar{X} = \frac{\sum f_i X_i}{N} = \sum p_i X_i.$$

The first formula is simpler, just the sum of the scores divided by the number of them. The second formula is more efficient when there are a large number of scores having only a few unique values. A formula for percentile points, and a formula for the median which is the 50th percentile point:

The general percentile point formula:

$$X_P = L + \left[\frac{P(N) - cf_b}{f_w} \right]$$

That same formula, adapted for the median by setting P to 0.50:

$$X_{50} = L + \left[\frac{.50\,\mathrm{N} - cf_b}{f_w} \right]$$

A1.2 Study Questions for Appendix 1

Central Tendency: Mean, Median, and Mode

Essay Questions

Q1 Explain the mean, median, and mode as physical quantities, i.e. what would they be on an actual physical distribution made of wooden blocks?

Q2 How are the mean, median, and mode related to one another in a positively skewed distribution? How are they related in a negatively skewed distribution?

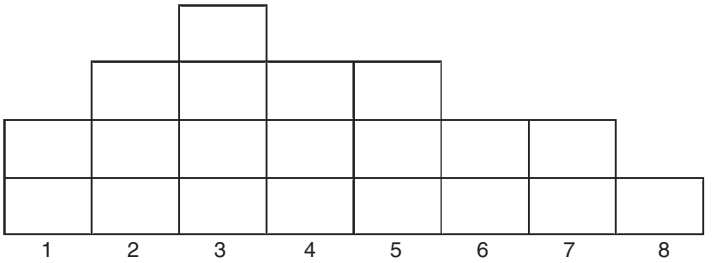
Q3 Which of the three central tendency measures is most stable, both in not being altered by local internal changes and also in not being overly altered by a few extreme values?

Q4 Which of the three measures of central tendency is most often used in statistical analyses? Why?

Q5 In what kinds of experimental settings would the median be the preferred measure and for what reasons?

Calculation Questions

Q6 Calculate the mean, median, and mode for the distribution shown in "stacked blocks" form using the formulas and concepts presented in Appendix 1.



Q7 As everyone familiar with the operation of see-saws (teeter-totters) at a children's park knows the balance of a plank with a fulcrum point in the middle depends on both the magnitude of the weights on the two sides and also on the distance of those weights from the fulcrum. The physical law is that the force of each weight source on the plank is the product of the weight (perhaps in grams) multiplied by the distance from the fulcrum (perhaps in centimeters). Suppose that the 22 blocks of the main distribution shown in Appendix 1 are each made of a copper alloy, each block having a volume of one cubic centimeter, with each block weighing 10 g. Using the physical equation that $F = W \times D$, where "F = force," "W = weight," and "D = distance-from-the-fulcrum," demonstrate that the force exerted on the plank to the right of the mean of 4.0 is equal to the force exerted on the plank to the left of that mean. In other words, demonstrate that the plank balances at the mean.

Q8 Now try another point along the plank, such as the median or the mode, and demonstrate the extent of imbalance for that point. That is, calculate the difference in force exerted on one side of the plank compared to the other. What do you learn from this?

Appendix 2

Calculation of Means, Variances, and Correlation Coefficients by the Algebra of Expectations

APPENDIX MENU

A2.1 Sample Mean as the Simplest Case of an Expected Value, 491
A2.2 Methods for Quantifying the Spread of a Distribution, 493
A2.2.1 Mean Absolute Deviation (MAD), 494
A2.2.2 Variance, 495
A2.2.3 Standard Deviation, 497
A2.2.4 Biased and Unbiased Variance, 498
A2.2.5 Standard Scores (*Z*-Scores), 499
A2.3 Methods for Quantifying Relatedness: Covariance and Correlation, 500
A2.3.1 Covariance, 500
A2.3.2 The Pearson Product-Moment Correlation Coefficient as the Expected Product of *Z*-Scores, 501
A2.4 Summary, 502
A2.5 Study Questions for Appendix 2, 503

This mathematical-concepts appendix investigates one of the most useful organizing concepts in the mathematics of data analysis, "the algebra of expectations." It will be illustrated as it is applied to the explanation of three elementary methods which together constitute the foundation of the art and science of data analysis, and also of psychometrics. First is the primary method for describing the central tendency of a distribution—the arithmetic mean. Second, we consider two closely related methods for describing the spread of a distribution—the variance and the companion statistic derived from it, the standard deviation. The standard deviation in turn leads to a particularly useful psychometric tool, the process of standardizing variables. Third, we consider the correlation coefficient and covariance, two conceptually and mathematically connected methods for describing data structure, that is, for describing how variables are related to one another.

A2.1 Sample Mean as the Simplest Case of an Expected Value

A small fictitious distribution with $N = 5$ scores is shown in Table A2.1. On the right side of that table, the average of the five scores is calculated. We refer to the average more formally as the "mean," which is short for the *arithmetic mean*. There are other types of means[1] but the arithmetic

1 The other two types of mean invented by the Pythagoreans are the geometric mean (useful in finance and anything with an exponential growth rate) and the harmonic mean (useful in averaging travel speeds, for example).

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

Table A2.1 Distribution of five *X* scores, with a mean equal to 6.

X	*N* = 5
9	
7	
6	$\overline{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$
5	
3	
$\sum X = 30$	

mean—the ordinary average—is by far the most familiar and the most useful. It is common, ubiquitous, and perhaps the most universally used of all statistical indices.

Most of us are quite familiar with this simple index and how to calculate it, but perhaps not so familiar with the formal notation used to express the calculation process, as shown in Table A2.1. The *X* with the bar above it, $\overline{X}$, is the usual symbol for the arithmetic mean of variable *X*. The capital Greek letter sigma (Σ) in the numerator of the equation of Table A2.1, followed by "*X*," represents the mathematical instruction to "sum the *X* values." Dividing that sum by *N*, the number of values, yields a mean. For the data of Table A2.1, the sum is 30, *N* is 5, and the mean is 6.

The arithmetic mean can also be referred to as the "expected value" of the scores. The reason may be obvious. Suppose the scores shown in Table A2.1 are scores of five persons on a particular neuropsychological test. If the score of an additional person from among the test takers was not yet known, the best guess would be 6, the mean. In other words, the mean gives the value that is expected, or most likely, for an unknown score. The mean is also an expected value in a more formal sense, in a *least squares* sense. That is to say, *the sum of squared deviations of a set of scores around the mean is less than around any other value*. The mean is therefore a summary statistic that fits the "least squares" criterion.

There happens to be a formula for obtaining the expected value that is more general, more difficult, and in important ways also more useful. This formula calculates the expected value as the sum of the products obtained by multiplying each score by its own probability, $E(X_i) = \sum p_i X_i$. This is numerically demonstrated below. Each of the five X_i values of Table A2.1 has a probability p_i of 1/5. Probability is the "relative frequency of occurrence" and each of the five values has a frequency of one, relative to five, the total of the frequencies. This process of summing the products of each p_i multiplied by its corresponding X_i yields a mean value of 6.

$$mean = \overline{X} = E(X_i) = \sum p_i X_i = \left(\frac{1}{5}\right)9 + \left(\frac{1}{5}\right)7 + \left(\frac{1}{5}\right)6 + \left(\frac{1}{5}\right)5 + \left(\frac{1}{5}\right)3 = \frac{30}{5} = 6$$

Notice that the result from this "probabilities-weighting" equation for the mean is mathematically equivalent to the result from the well-known sum-of-the-scores-divided-by-*N* method used in Table A2.1.

Although this general-purpose formula for an expected value is needlessly complex for merely obtaining an average, it has two advantages, a major one and a minor one. The minor advantage is one of efficiency of some types of data distributions. For example, when a data set has a large number of observations but a small number of unique values, such as the distribution shown in

Table A2.2 Frequency distribution of X scores for a large sample.

X_i	f_i	p_i
4	12	12/60
3	21	21/60
2	18	18/60
1	9	9/60
$N = 60$		

Table A2.2, which has an N of 60 but only four unique values, this method of calculating the mean is economical. Each unique value appears only once in the calculations, while the relative number of occurrences is reflected in the probabilities. The values of 4, 3, 2, and 1, respectively, have frequencies of occurrence of 12, 21, 18, and 9. The probabilities, that is, the relative frequencies, are derived by dividing each frequency by 60, the total number of scores. The mean can then be readily and efficiently calculated using the expected-values probabilities-weighting equation. Even though there are 60 observations, we only need to sum four weighted values to get the mean—a textbook example of efficiency.

$$mean = \overline{X} = E(X_i) = \sum p_i X = \left(\frac{12}{60}\right)4 + \left(\frac{21}{60}\right)3 + \left(\frac{18}{60}\right)2 + \left(\frac{9}{60}\right)1 = \frac{156}{60} = 2.6$$

The second advantage, and by far the most important, is pedagogical. The expected-values method provides a simplified conceptual/mathematical structure for calculating a number of additional statistical indices such as the *variance* $= E\left(x_i^2\right) = E\left(X_i - \overline{X}\right)^2$, which is the expected value of squared deviation scores; the *covariance* $= E(x_i y_i)$, which is the expected value of the product of deviation scores across two variables; and the *correlation coefficient* $= E(Z_x Z_y)$, which is the expected value of the product of Z-scores[2] across two variables. The expected-values approach reveals the conceptual unity and the similarity across a wide variety of the most useful and fundamental statistical indices.

These additional statistical indices, introduced and explained below, increase our analytic capabilities from merely being able to find the central tendency of a distribution, to accomplishing such things as quantifying the spread or "variation" of scores within a single distribution; or quantifying the predictive relationship between two or more such variables, and thus being able to quantify the holistic structure of a large number of variables. The first illustration of the broad applicability of expected value formulas will be to compare three ways of measuring spread.

A2.2 Methods for Quantifying the Spread of a Distribution

In this section, three approaches to the quantification of the spread of a distribution of data are demonstrated and compared. The first, mean absolute deviation (MAD), is merely a pedagogical device. It is not used in actual practice but is instrumental in providing a clear understanding of the other two measures of spread, variance (S^2) and the standard deviation (S). These two are

2 Z-scores, also called standard scores, are explained in Section A2.5.

absolutely central and universally used. They are among the half dozen most important building blocks in statistical science. In Section A2.2.5, the standard deviation will be used to demonstrate the process of the standardization of variables using *Z*-scores.

A2.2.1 Mean Absolute Deviation (MAD)

The most direct and obvious way to quantify spread is to simply calculate the average distance of scores from their mean. This is called the *mean deviation*. Formally, it may be defined as the *expected value of the deviation scores*, where deviation scores are found by subtracting the mean from each raw score.

$$\textit{deviation score}_i = x_i = X_i - \overline{X}$$

The deviation score corresponding to the first raw score for the small dataset introduced in Table A2.1 is calculated as $x_1 = X_1 - \overline{X} = 9 - 6 = 3$. Similarly, the deviation score corresponding to the fourth score is calculated as $x_4 = X_4 - \overline{X} = 5 - 6 = -1$. For the five Table A2.1 raw scores of 9, 7, 6, 5, and 3, the corresponding deviation scores are therefore 3, 1, 0, −1, −3. The raw scores are symbolized as X_i, with a capital X, and the deviation scores are symbolized as x_i, with a small x. The subscript "i" indicates that X is a variable, meaning that it can take on a number of differing values, as opposed to being a constant. For example, the first value in the distribution is written as X_1 with its deviation score written as x_1, the second as X_2 with its deviation score as x_2, and so on. The general term would be X_i where the subscript "i" represents abstractly any one of the subscripts, without specifying a particular one.

It should be noted, however, that there is a problem with this first suggested definition of the mean deviation statistic. The expected value, or average, of the deviation scores is found to be zero, and this will be true for all distributions.

$$E(x_i) = \sum p_i x_i = \left(\frac{1}{5}\right)3 + \left(\frac{1}{5}\right)1 + \left(\frac{1}{5}\right)0 + \left(\frac{1}{5}\right)(-1) + \left(\frac{1}{5}\right)(-3) = \frac{0}{5} = 0$$

We must redefine the mean deviation statistic as the expected value of *the absolute* values of the deviation scores, the MAD statistic, MAD for short.

$$MAD_x = E(|x_i|) = \sum p_i |x_i| = \left(\frac{1}{5}\right)|3| + \left(\frac{1}{5}\right)|1| + \left(\frac{1}{5}\right)|0| + \left(\frac{1}{5}\right)|-1| + \left(\frac{1}{5}\right)|-3| = \frac{8}{5} = 1.6$$

The sum of the five deviation score values (when taking absolute values), as shown in the foregoing equation and also in Table A2.3, is found to be 8. This sum of absolute values of deviations is

Table A2.3 MAD calculation from deviation scores, x_i.

x_i	$N = 5$
3	
1	
0	$\overline{X} = \dfrac{\Sigma \mid x_i \mid}{N} = \dfrac{8}{5} = 1.6$
−1	
−3	
$\sum\|x_i\| = 8$	

Table A2.4 Distribution of five *Y* scores, with a mean equal to 12.

Y_i	$N = 5$
18	
14	
12	$\overline{Y} = \frac{\Sigma Y_i}{N} = \frac{60}{5} = 12$
10	
6	
$\sum Y_i = 60$	

divided by *N*, which is 5 in this example. So, the mean of the absolute values of the deviations is 1.6. This is an easy index to calculate, and easy to interpret. We can say that, on average, these five scores deviate 1.6 units from their mean in either a positive or a negative direction.

The probability-based expected-values equation for MAD just given is mathematically equivalent to the simpler algebraic expression $MAD_x = \frac{\sum|x_i|}{N}$ shown in Table A2.3. The numerator in this equation is the sum of the absolute values of the deviation scores, 8. Dividing this sum by *N*, the sample size, is equivalent to multiplying each absolute deviation score by its probability (which is $\frac{1}{N}$ for each of them) and then summing them. One can either multiply each score by $\frac{1}{N}$ individually and then sum them, or first sum them all and then divide the sum by *N*. The two have the same effect, and either will produce the correct value of the MAD.

Consider now the distribution of five scores for variable *Y* shown in Table A2.4. This small distribution is comparable to variable *X* in Table A2.1. It also consists of five values but each value is exactly twice as large as the corresponding five values of *X*. From the calculations in Table A2.4, we see the mean of this distribution is 12, exactly twice the value of the mean of *X*. The deviation scores for distribution *Y* (6, 2, 0, −2, and −6) are also each twice as large as those for *X*. As shown below, the statistic MAD for *Y* is 3.2, also exactly twice the size of the mean deviation statistic for *X*.

$$MAD_y = E(|y_i|)$$
$$= \sum p_i|y_i| = \left(\frac{1}{5}\right)|6| + \left(\frac{1}{5}\right)|2| + \left(\frac{1}{5}\right)|0| + \left(\frac{1}{5}\right)|-2| + \left(\frac{1}{5}\right)|-6| = \frac{16}{5} = 3.2$$

Although the mean deviation is easy to calculate and has a clear interpretation, it is almost never used in actual statistical practice. The two measures of spread that *are* in common usage for quantitative data are the variance (S^2) and the standard deviation (S). In what follows, we will present these two statistics, show how they are calculated, explain the relative properties of each, and why both are preferred over the mean deviation statistic.

A2.2.2 Variance

Even though the MAD statistic is not used in common practice, it is pedagogically useful. Understanding the MAD statistic makes it easier to comprehend and remember the conceptual basis of variance, as shown in its simplest algebraic equation, $S_x^2 = \frac{\sum x_i^2}{N}$. Instead of using absolute values to

remove the minus signs found on the deviation scores (approximately half of them), in the variance formula, the deviation scores are made positive by squaring. Other than that, the two equations are the same. Therefore, whereas the mean deviation is the expected value of the *absolute values* of the deviation scores, variance is the expected value of the *squared* deviation scores, as shown below for both distribution X and also for distribution Y. The symbol for variance is "S-squared," S^2, to remind us that variance is expressed in squared units.

$$S_x^2 = E(x_i^2) = \sum p_i x_i^2 = \left(\frac{1}{5}\right)3^2 + \left(\frac{1}{5}\right)1^2 + \left(\frac{1}{5}\right)0^2 + \left(\frac{1}{5}\right)(-1)^2 + \left(\frac{1}{5}\right)(-3)^2 = \frac{20}{5} = 4.0$$

$$S_y^2 = E(y_i^2) = \sum p_i y_i^2 = \left(\frac{1}{5}\right)6^2 + \left(\frac{1}{5}\right)2^2 + \left(\frac{1}{5}\right)0^2 + \left(\frac{1}{5}\right)(-2)^2 + \left(\frac{1}{5}\right)(-6)^2 = \frac{80}{5} = 16.0$$

Notice the effect that doubling X to create Y has on the variance of Y. The variance of Y becomes four times as large as the variance of X. In general, if you transform a variable by multiplying it by any constant, c, it will multiply the variance of that variable by the square of that constant, c^2.

$$E\left[(c \cdot x_i)^2\right] = c^2 E(x_i^2)$$

For example, when $c = 2$, thus doubling the values in the transformed variable, the variance of the new variable is four times the variance of the old.

$$S_{new}^2 = c^2 S_{old}^2 = 2^2(4) = 4 \times 4 = 16$$

Had we multiplied the X scores by 3 to get a new variable, the variance of that new variable would have been 36.

$$S_{new}^2 = c^2 S_{old}^2 = 3^2(4) = 9 \times 4 = 36$$

As mentioned above, the mean is a "least squares" statistic. That is, the sum of the squared deviations will be less around the mean than around any other value. It follows that since variance *is* the sum of squared deviations divided by N, that is $S_x^2 = \frac{\sum x_i^2}{N}$, *the variance of scores around the mean will be less than the variance around any other value*. This can be demonstrated. The sum of squared deviations around the mean of 6 for the X scores is: $\sum x_i^2 = 3^2 + 1^2 + 0^2 + (-1)^2 + (-3)^2 = 20$, with the variance being $20/5 = 4$. However, the sum of squared deviations around a value exactly 1 greater than the mean of 6, a value of 7, is found to be quite a bit larger than 20.

$$\sum x_i^2 = (9-7)^2 + (7-7)^2 + (6-7)^2 + (5-7)^2 + (3-7)^2$$

$$= 2^2 + 0^2 + (-1)^2 + (-2)^2 + (-4)^2 = 25$$

The sum of squared deviations around a value of 5, exactly one less than the mean of 6, is also found to be 25, as the reader can verify. In general, the sum of squared deviations around any value greater than the mean, or less than the mean, is greater than the sum of squared deviations around the mean itself. Therefore, the mean is a least squares statistic, in other words, a statistic defined by the least squares criterion. As such, it is our first example of a least squares statistic, that is, a statistic

designed such that the sum of squared deviations is a minimum. We shall also encounter additional least squares statistics in regression analysis and other methods throughout the book.

Even though squaring, as just shown, does effectively remove the negative values from deviation scores, it is much more than just an adjustment factor for convenience. An important reason variance is preferred over MAD, as the index of choice for quantifying spread, is that variance can be partitioned into additive parts. Because it is expressed in squared units, variance can be subdivided into additive pieces, such as that part of the variance which is predictable, $S^2_{predicted} = 64$ for example; and that part which is not predictable, $S^2_{error} = 36$. These two parts are additive and will sum to the total variance, $S^2_{total} = 100$. This will be fully demonstrated and explained in Appendix 4, Bivariate Regression.

Variance is consistent with the principles of Euclidean geometry, in a way that MAD is not. As expressed in the Pythagorean Theorem, the square of the hypotenuse of a right triangle is equal to the sum of the squares of the other two sides, $c^2 = a^2 + b^2$. The magnitudes of the sides do not sum to the hypotenuse, but the *squares* of the sides do sum to the *square* of the hypotenuse. Squared indices are therefore Pythagorean, additive, and capable of being partitioned. This variance-partitioning property will be of much value to us in statistical and psychometric work. Statistical methods are firmly rooted in Euclidean geometry, as is matrix algebra, and variance is the central Pythagorean index upon which much of statistical science is based.

A2.2.3 Standard Deviation

Standard deviation is a derivative index of spread—it is derived from the variance. It is, in fact, the square root of the variance, $S = \sqrt{S^2}$. Variance is based upon the *squared magnitude* of the variable, a summary descriptive index of the spread of the distribution in squared units; and standard deviation is the *magnitude* of the variable in unsquared units. For variable X given above, the standard deviation is:

$$S_x = \sqrt{S_x^2} = \sqrt{4} = 2$$

Similarly, the standard deviation for Y is:

$$S_y = \sqrt{S_y^2} = \sqrt{16} = 4$$

Even though squared units are useful for partitioning information, unsquared units still have an advantage in terms of interpretability. Suppose, for example, that our variables X and Y shown above were scores on a computerized cognitive task along the lines of the Digit-Symbol Substitution Test (a useful neuropsychological test where patients are asked to match a symbol to its assigned digit using a lookup table). The score is the number of correct matches within a fixed time limit. Suppose also there are two versions of the cognitive task, one more difficult and one easier, and that variable X above represents the five subjects' scores on the more difficult test since they score fewer correct matches. Variable Y is their scores on the easier test. We could give a direct interpretation of both the means and also of MAD for these two variables. The means ($\overline{X} = 6$ and $\overline{Y} = 12$) tell us that on the Y task (the easier one), the subjects had twice as many correct matches within the time limit as on the X task. Also, the MADs ($MAD_x = 1.6$ and $MAD_y = 3.2$) indicate that their scores on the Y task are twice as spread out as on the X task.

With the two variances ($S_x^2 = 4$ and $S_y^2 = 16$), however, the comparisons are not so easily interpreted. It does not make sense in terms of the meaning of the variables to say that the easy task has four times as many "average *correct-squared* matches" as the difficult task. Variance makes geometric sense, but it does not have the interpretive transparency found in the MAD statistic. Unsquared units are more directly interpreted than squared units. The solution to this dilemma is the *standard deviation*, an index which is interpretable in the unsquared-units-of-measure sense, and yet also makes geometric sense. Mathematically and geometrically, the standard deviation is the *magnitude* of a vector. That is, variables such as X and Y can be represented geometrically as vectors in a multidimensional space, as is shown in *Appendix 3*, *Matrix Multiplication*. The variance of X is the squared length (squared magnitude) of the vector representing X, and the standard deviation is the *magnitude* (that is, the actual length) of that vector.

The standard deviation index still does not have the simple and direct "average-of-deviation-scores" interpretation that characterizes the MAD, but it does have the advantage of being expressed in unsquared units of "number of correctly matched" digit-symbol substitutions. We also see that the standard deviations maintain the correct ratio for our fictitious neuropsychological test data, that of exactly twice as much spread in the Y scores ($S_y = 4$) as in the X scores ($S_x = 2$).

The variance is the mean of the squared deviations of scores from their own arithmetic mean. For this reason, variance is often referred to as an MS, that is, a "mean square." Similarly the standard deviation, as the square root of the variance, is often referred to as the "root mean square" or RMS. The term *RMS error*, which appears in Appendices 4 and 5 and in several chapters, refers to the standard deviation of error scores in a regression analysis.

A2.2.4 Biased and Unbiased Variance

The discussion of variance and standard deviation is not complete without an explanation of the difference between biased and unbiased variance. The formula for variance and the formula for standard deviation presented above work fine when applied to populations, or when simply used to describe a sample dataset. However, in statistical and psychometric work, the focus is almost always inferential. That is, almost always the task is to use the sample data to draw conclusions regarding the population from which it came. The word "sample" implies testing a part to infer the nature of the whole, the entire population.

Strange as it might seem, it is demonstrably true, that a variance calculated by the formula $S^2 = \frac{\sum x^2}{N}$ gives a *biased* estimate of the true population variance, and that bias will be larger the smaller the sample size is. The biased sample variance will systematically underestimate the true population variance by one Nth, where N is the number of scores entering into the calculation of the mean and the variance. The formula for an *unbiased* sample variance, $\hat{s}^2 = \frac{\sum x^2}{N-1}$, is very similar to the formula for the biased one. We simply divide the sum of squared deviations by $N - 1$ rather than by N. This $N - 1$ in the denominator is referred to as the "degrees of freedom."

Notice that a squared lowercase s is used as the symbol for an unbiased variance and that the squared s has a hat over it. The hat signifies that the estimate is in some sense statistically optimal. In this case, its optimality is that it is unbiased, i.e. does not have a systematic tendency to underestimate nor to overestimate the true population variance. By contrast, a squared uppercase S is used as a symbol for a biased sample variance. The formula for the variance of a population is $\sigma^2 = \frac{\sum x^2}{N_p}$ where the squared Greek letter sigma σ^2 signifies the population variance, and N_p signifies the number of cases in the population. Suppose that the actual population variance for our

example X variable is 5, that is, $\sigma^2 = \frac{\sum x^2}{Np} = 5$. It would follow then that in many sample estimates of that population variance, the expected value of estimates using the degrees of freedom formula, the expected value would be 5, $E(\hat{s}^2) = 5$, whereas if our sample size were $N = 10$, the expected value of estimates using the biased formula would be $E(S^2) = \frac{\sum x^2}{N} = 4.50$. In other words, the biased formula underestimates the true population value by one Nth of its value, which is one-tenth in this case: $\left(\frac{1}{N}\right)\sigma^2 = \left(\frac{1}{10}\right)5 = 0.50$. If the sample size were 100, we would in the long run underestimate σ^2 by one 100th. The expected value of S^2 would be $E(S^2) = 4.95$.[3]

A2.2.5 Standard Scores (*Z*-Scores)

A standard score, or Z-score as it is often called, is the *signed number of standard deviation units that a score is above or below its mean*. Its purpose is to make scores from distributions that are not commensurate with one another, that is, distributions that differ in means and in standard deviations, directly comparable in relative magnitudes. For example, suppose you are person C in the table of bivariate values in Table A2.4, and you have a score of 6 on task X and a score of 10 on task Y. At first look, it would appear that you did better on task Y, with a score of 10, than on task X with a score of 6. But your score of 6 on task X is actually equivalent to the mean score, while your score of 10 on task Y, the easier task, is two units below the task Y mean of 12. Since the standard deviation of task Y is 4, your deviation score of −2 is one-half of a standard deviation unit below the Y mean, a Z-score of −0.5, while your score on task X has zero deviation from its mean, a Z-score of 0.0. So, when we turn the X and Y scores into Z_x and Z_y scores, which tell you how many standard deviation units you are above or below the mean, as shown in Table A2.5, you see accurately and comparatively your relative standing on each task.

The equation for creating standard scores is simple. A standard score, that is, a Z-score, is just the deviation score divided by its standard deviation: $Z_x = \frac{x_i}{\hat{S}_x}$. Using this equation, we calculate the Z-scores for person C on task X and task Y.

Table A2.5 Bivariate scores of five persons on two variables.

Person	*X*	*Y*
A	9	14
B	7	18
C	6	10
D	5	12
E	3	6

3 This principle is demonstrated algebraically in section 13, at the end of Chapter 1.

$$Z_x = \frac{x_i}{\hat{s}_x} = \frac{X_i - \overline{X}}{\hat{s}_x} = \frac{6-6}{2} = \frac{0}{2} = 0.0$$

$$Z_y = \frac{y_i}{S_y} = \frac{Y_i - \overline{Y}}{S_y} = \frac{10-12}{4} = \frac{-2}{4} = -0.5$$

Similar calculations are used to obtain *Z*-scores on each of the five *X* scores and the five *Y* scores as recorded in Table A2.6, found in Subsection §A2.3.2 of this appendix. We can see from this table that two of the persons (A and C) did better, relatively, on task *X* than on task *Y*, two of them (B and D) did better on task *Y* than on task *X*, and one (person E) did equally well on the two tasks. The relative magnitudes of these comparative values become obvious when we examine the table of *Z*-scores, but would not be obvious from the raw scores of the five persons on the two tasks. "Standard scores" is, therefore, a good name for *Z*-scores, in that they give us a standard template for comparing scores across distributions which are not commensurate, that is, which differ in their central tendency (means) and/or in their spread (standard deviations). As we progress through the book, we will find that standard scores in combination with the bell-shaped curve, which when combined are referred to as the "standard normal curve," are particularly useful for interpreting data.

A2.3 Methods for Quantifying Relatedness: Covariance and Correlation

This final section introduces methods for dealing with multiple variables, that is, methods that can describe multivariate data structure and the relatedness of variables to one another. It can be done with covariances, or it can be done with correlation coefficients.

A2.3.1 Covariance

Covariance, as the name implies, is related to variance. However, whereas variance is a spread index for a single variable, covariance is a measure of the extent to which two variables vary in concert with one another, that is, the extent to which one variable is predictable from the other.

For example, we may wish to assess how well one can predict the scores of each of five persons on task *Y*, the easier cognitive task, from their scores on the more difficult one, task *X*. Covariance and related indices are intended to answer this question of predictability, that is, how predictable variable *Y* is from variable *X*. Paired scores for task *X* and task *Y* for each of the five persons (A, B, C, D, and E) are shown in Table A2.5. This is called a bivariate distribution because each person has two scores that belong to her or him, one on task *X* and one on task *Y*. The two scores are linked, in that on each row both scores belong to the same person.

Notice that the *Y* scores are not in descending order here as they were when scores were displayed for each task separately in the tables above. That is because the five persons' scores, in our fictitious example, do not have the same ordinal properties on task *Y* as they do on task *X*. Although person A has the highest score on task *X*, she has the second highest score on task *Y*. Person B, on the other hand, has the second-highest score on task *X*, but the highest on task *Y*, and so forth. The data pairings *are* what they *empirically are* and those links must not be changed. This is because the ordinal positions of the five persons' scores on task *Y* in relation to the ordinal positions of their scores on task *X* contain the information regarding how related (that is, how 'co-related') the two variables are. You cannot arbitrarily alter the empirical pairings of scores without destroying the information about the correlation/covariance of the two variables.

The simplest algebraic equation for covariance is $S_{xy} = \frac{\sum x_i y_i}{N}$. You will notice that this looks very much like the algebraic formula for variance given above, except that the numerator is the sum of *products* (also referred to as "product sum") of the deviation scores x_i and y_i, whereas the numerator for variance was the sum of *squares* of the x_i deviation scores. Therefore, whereas variance is the expected value of the *squared* deviation scores on a single variable, covariance is the *expected value of the product of the deviation scores across two variables*, X and Y, as shown. The deviation scores for task X are 3, 1, 0, −1, −3, appearing in descending order. On the other hand, the deviation scores for task Y are 2, 6, −2, 0, −6, and depart somewhat from the descending order of X, indicating that the two variables are not perfectly correlated. The calculation of covariance from the expected-values-of-deviation-scores equation is demonstrated below.

$$S_{xy} = E(x_i y_i) = \sum (p_i)(x_i y_i)$$

$$= \left(\frac{1}{5}\right)(3)(2) + \left(\frac{1}{5}\right)(1)(6) + \left(\frac{1}{5}\right)(0)(-2) + \left(\frac{1}{5}\right)(-1)(0) + \left(\frac{1}{5}\right)(-3)(-6) = \frac{30}{5} = 6.0$$

The Pearson product-moment correlation coefficient (PPMCC) is defined as *standardized covariance*. We standardize the covariance by dividing it by the standard deviation of X and the standard deviation of Y.

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{6.0}{(2)(4)} = \frac{6.0}{8} = 0.75$$

From this calculation, we see that task X and task Y have a correlation coefficient of 0.75, and are therefore somewhat positively related.[4] In Appendix 4, we present additional information about how to interpret the value of the PPMCC.

A2.3.2 The Pearson Product-Moment Correlation Coefficient as the Expected Product of *Z*-Scores

We saw in the covariance section above that the PPMCC can be obtained from the covariance by dividing the covariance by its two standard deviations, or in other words, by standardizing the covariance. The PPMCC, symbolized as r, can also be obtained by an alternative process, that is, by calculating a covariance on already obtained standardized scores (Z-scores). In other words, we can do the standardizing either before or after the covariance-calculating process.

Using the second approach, that is the approach of first calculating Z-scores and then calculating a covariance of them, the correlation coefficient r can be defined as the *expected value of the products of standard scores*. The Z-score values corresponding to the raw score values in Table A2.5 are given in Table A2.6. The calculation of the PPMCC using the expected-values equation is demonstrated below.

4 The correlation coefficient is an index of the linear match between the ordinal and interval properties of scores on the two variables. A perfect match (that is, a perfect positive linear relationship between the two), has a correlation coefficient of $r = 1.00$. A perfect negative (inversely related) linear relationship has a correlation coefficient of $r = -1.00$. When the two variables are unrelated (that is, orthogonal or perpendicular), the correlation coefficient is $r = 0.00$.

Table A2.6 Bivariate *Z*-scores corresponding to the bivariate raw scores of Table A2.5.

Person	Z_X	Z_Y
A	1.5	0.5
B	0.5	1.5
C	0	−0.5
D	−0.5	0
E	−1.5	−1.5

$$r_{xy} = E(Z_x Z_y) = \sum p_i(Z_x Z_y) = \left(\frac{1}{5}\right)(1.5)(0.5) + \left(\frac{1}{5}\right)(0.5)(1.5) + \left(\frac{1}{5}\right)(0)(-0.5) + \left(\frac{1}{5}\right)(-0.5)(0) + \left(\frac{1}{5}\right)(-1.5)(-1.5) = \frac{3.75}{5} = 0.75$$

A2.4 Summary

The algebra of expectations shows us how similar these six statistical indices are to one another. The *mean* is the expected value of raw scores. The MAD is the expected value of absolute deviation scores. The *variance* is the expected value of squared deviation scores. (Notice in the table how one can convert a biased sample variance S_x^2 to an unbiased sample variance.) The biased *standard deviation* is the square root of the expected value of squared deviation scores, in other words, the square root of the biased variance. To get an unbiased standard deviation, simply take the square root of the unbiased variance. The *covariance* is the expected value of products of deviation scores across two variables, and the PPMCC is the expected value of the products of Z_x and Z_y across two variables.

Mean	$\overline{X} = E(X_i)$	and	$\overline{X} = \frac{\sum X_i}{N}$
Mean absolute deviation	$MAD = E(\lvert x_i \rvert)$	and	$MAD = \frac{\sum \lvert x_i \rvert}{N}$
Variance (biased)	$S_x^2 = E(x_i^2)$	and	$S_x^2 = \frac{\sum x_i^2}{N}$
Unbiased variance	$\hat{s}_x^2 = \left(\frac{N}{N-1}\right) S_x^2$	and	$\hat{s}_x^2 = \frac{\sum x_i^2}{N-1}$
Standard deviation	$S_x = \sqrt{E(x_i^2)}$	and	$S_x = \sqrt{\frac{\sum x_i^2}{N}}$
Unbiased standard deviation	$\hat{s}_x = \sqrt{\hat{s}_x^2}$	and	$\hat{s}_x = \sqrt{\frac{\sum x_i^2}{N-1}}$
Biased covariance	$S_{xy} = E(x_i y_i)$	and	$S_{xy} = \frac{\sum x_i y_i}{N}$
Unbiased covariance	$\hat{s}_{xy} = \left(\frac{N}{N-1}\right) S_{xy}$	and	$\hat{s}_{xy} = \frac{\sum x_i y_i}{N-1}$
Correlation coefficient	$r = E(Z_x Z_y)$	and	$r = \frac{\hat{s}_{xy}}{\hat{s}_x \hat{s}_y}$

A2.5 Study Questions for Appendix 2

Calculation of Means, Variances, and Correlation Coefficients by the Algebra of Expectations

Essay Questions

1 Explain what is meant by the statement that the mean is a "least squares" statistic.

2 Explain how the "expected-values" formula for a mean ($\overline{X} = E(X_i) = \sum p_i X_i$) has advantages over the common formula for a mean ($\overline{X} = \frac{\sum X}{N}$) when dealing with very large distributions containing a relatively small number of unique values.

3 How is the variance statistic (a biased variance, that is) similar to the MAD statistic? What is the only difference between them? Why is MAD never used?

4 How is the preference for the variance statistic over the MAD statistic related to the Pythagorean Theorem? In other words, what is the reason for and the advantage of incorporating the squaring of values into our preferred index of spread?

5 What are the advantages of variance as a measure of spread over the standard deviation? What are the advantages of the standard deviation as a measure of spread over variance? Under what conditions, or for what uses, is each the preferred measure?

6 How are the three statistical indices—variance, covariance, and the Pearson Product Moment Correlation Coefficient (PPMCC)—related to one another, both conceptually and mathematically?

7 How are standard scores useful? How do they make variables with noncommensurate measurement scales (that is, incomparable to one another) comparable?

8 Write the three formulas that express each of the following definitions of the Pearson Product Moment Correlation Coefficient (PPMCC): (i) the PPMCC as "standardized covariance," (ii) the PPMCC as the "covariance of standard scores," and (iii) the PPMCC as the "expected value of the products of standard scores."

Calculation Questions

1 Consider the following small and simple dataset consisting of $N = 7$ bivariate values.

X	*Y*
12	7
11	10
9	5
9	2
8	5
8	5
6	1

Calculate the unbiased variance of variable X by the following simple "definitional" formula.

$$s^2 = \frac{\sum x_i^2}{N-1}$$

2 Now calculate the unbiased variance of variable X by the following more precise "computational" formula and verify that the two formulas give the same result. What is the standard deviation for this variable?

3 Finally, calculate the biased variance of variable X by the following "definitional" formula with N in the denominator rather than $N-1$, the "degrees of freedom." Which is smaller, the unbiased variance or the biased one? How much smaller is it? Does your result verify the statement that "biased sample variance underestimates the true population variance by one Nth," where N is the sample size of the calculation?

$$S^2 = \frac{\sum x_i^2}{N}$$

4 Now calculate the unbiased variance of variable Y above by the following simple "definitional" formula:

$$s^2 = \frac{\sum y_i^2}{N-1}$$

5 Calculate the unbiased covariance between variable X and variable Y using the following simple "definitional" formula.

$$s_{xy} = \frac{\sum x_i y_i}{N-1}$$

6 Calculate the correlation coefficient on this dataset using the variances of X and Y and the covariance from questions 1, 4, and 5.

7 Suppose you are an officer in charge of the Alpha Barracks, which houses 120 first-, second-, third-, and fourth-year cadets. A new group of 120 cadets will arrive next week and you now have the frequency distribution of how many are in year 1, how many in year 2, how many in year 3, and how many in year 4, as shown in the frequency distribution below. You sum up the four frequencies and verify that there are indeed 120 cadets. You then divide each frequency by 120 to obtain the probability of cadets in Alpha Barracks being in each of the four years. Use the "probabilities-weighting" expected-values formula given at the top of page 2 to obtain the average number of years your cadets have been in the program.

$$\overline{X} = E(X_i) = \sum p_i X$$

X	f	p
4	18	0.150
3	45	0.375
2	42	0.350
1	15	0.125

8 Now use the expected-values "probabilities-weighting" equation for a biased variance given at the top of page 4 in this appendix to obtain the biased variance of this distribution. As you may notice, these formulas save a substantial amount of time in calculating means and variances on a distribution that has a small number of unique values (four in this case) and a relatively large sample size (120 in this case).

$$S_x^2 = E\left(x_i^2\right) = \sum p_i x_i^2$$

9 Having calculated a biased variance in question 8, now convert it into an unbiased variance using the formula for conversion given in the fourth row of the table of formulas at the end of this appendix:

$$s_x^2 = \left(\frac{N}{N-1}\right) S_x^2$$

Appendix 3

Matrix Multiplication

APPENDIX MENU

A3.1 Matrices, Vectors, and Scalars, 506
A3.2 Instructions for Doing Matrix Multiplication Using an Excel Spreadsheet, 508
A3.3 Calculating Variance from Raw Data by Three Matrix Methods, 509
A3.4 Calculating the 3 × 3 Covariance Matrix Using the Hybrid Formula, 511
A3.5 Matrix Multiplication and the Equation for Probabilities-Weighting, 513
A3.6 Summary, 516
A3.7 Study Questions for Appendix 3, 516

In his 1896 paper, "Mathematical Contributions to the Theory of Evolution: Regression, Heredity, and Panmixia," Karl Pearson was ready to credit Bravais (1846) with being the first to envision the concept of a correlation coefficient as an index of relatedness. However, in 1920, with a more careful reading of Bravais, he reversed his earlier position and accorded that honor to Galton. Pearson (1920) indicated that although Bravais is probably the originator of the concept of a "product sum," a central concept used by Pearson in his development of the product moment correlation coefficient, there was no recognition in Bravais's writing of its usefulness as an index of relatedness. It appears there was no recognition of even the need for such an index. Pearson concluded that Galton (1888) first recognized that need and took the first steps toward a realization.

Bravais's "product sum" is a major part of statistical science as we know it today. It is a crucial and central concept, not only in the emergence of statistical theory and methods, but also in that branch of mathematics known as linear algebra, and in particular, in the process of matrix multiplication. In Appendix 2, where the algebra of expectations was introduced, it was demonstrated that the computational engine for covariance (and also variance for that matter, since *variance is a special case of covariance*) is the sum of products, ΣXY, or as Bravais referred to it, the product sum. If the products that we sum are the products of each X value times itself, ΣX^2, we have the numerator for variance. If the products that we sum consist of an X value times a Y value, ΣXY, we have the numerator for covariance. Matrix multiplication is also firmly based on the concept of the product sum. The first thing to be learned about matrix multiplication is that it consists entirely of "product sums" of matrix rows and columns in varying combinations with one another.

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

A3.1 Matrices, Vectors, and Scalars

Shown below are three examples of matrices (labeled $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$) at the top of page 506 followed by three examples of vectors (labeled $\boldsymbol{d}$, $\boldsymbol{e}$, and $\boldsymbol{f}'$). Vectors are a special case of matrices. A vector is a matrix that is one-dimensional, consisting of only a single row or a single column. The dimensions of matrices are specified as $r \times c$, where r is the number of rows and c is the number of columns. Matrix $\boldsymbol{A}$ is a 3×2 matrix since it has 3 rows and 2 columns. Matrix $\boldsymbol{B}$ is a 2×4 matrix having 2 rows and 4 columns, and $\boldsymbol{C}$ is a 4×6 matrix. The next, $\boldsymbol{d}$, is a 3×1 vector, a column vector; and $\boldsymbol{e}$ is a 2×1 column vector. Vector $\boldsymbol{f}'$ is a 1×3 row vector. A 1×1 matrix is just a single numerical value, such as $m = 7$, and is referred to as a "scalar."

$$\boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 1 & 4 \\ 6 & 2 \end{bmatrix} \quad \boldsymbol{B} = \begin{bmatrix} 5 & 1 & 7 & 2 \\ 3 & 3 & 5 & 4 \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} 1 & 5 & 7 & 2 & 1 & 3 \\ 4 & 6 & 1 & 5 & 2 & 4 \\ 3 & 1 & 7 & 8 & 6 & 2 \\ 4 & 2 & 3 & 5 & 4 & 2 \end{bmatrix}$$

$$\boldsymbol{d} = \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix} \quad \boldsymbol{e} = \begin{bmatrix} 7 \\ 5 \end{bmatrix} \quad \boldsymbol{f}' = [5 \;\; 2 \;\; 7]$$

Notice that it is customary for the algebraic symbols for matrices, such as $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$, to be in bold upper-case letters, and for vector symbols such as $\boldsymbol{d}$, $\boldsymbol{e}$, and $\boldsymbol{f}'$ to be in bold lower-case letters. The bolding indicates that $\boldsymbol{C}$, for example, is not just an ordinary scalar value, but a matrix containing a structural arrangement of multiple values.

Observe that the row vector $\boldsymbol{f}'$ has an apostrophe after the bold $\boldsymbol{f}$. This is read as "$\boldsymbol{f}$ transpose." The transpose of a matrix, or of a vector, is obtained by interchanging rows and columns. Consider the following two vectors and two matrices.

$$\boldsymbol{f} = \begin{bmatrix} 5 \\ 2 \\ 7 \end{bmatrix} \quad \boldsymbol{f}' = [5 \;\; 2 \;\; 7] \quad \boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 1 & 4 \\ 6 & 2 \end{bmatrix} \quad \boldsymbol{A}' = \begin{bmatrix} 3 & 1 & 6 \\ 2 & 4 & 2 \end{bmatrix}$$

The first vector, $\boldsymbol{f}$, is a 3×1 column vector.[1] When we transpose vector $\boldsymbol{f}$, that is, when we transform it from a column vector to a row vector, it becomes the 1×3 row vector $\boldsymbol{f}'$. Similarly, the matrix $\boldsymbol{A}$ is a 3×2 matrix, but when it is transposed to become matrix $\boldsymbol{A}'$, its rows and columns are interchanged, and it becomes a 2×3 matrix. Row 1 of matrix $\boldsymbol{A}$ consisting of a 3 followed by a 2, becomes column 1 of matrix $\boldsymbol{A}'$. Row 2 of matrix $\boldsymbol{A}$ (1 followed by 4) becomes column 2 of matrix $\boldsymbol{A}'$, and row 3 of matrix $\boldsymbol{A}$ (6 followed by 2) becomes column 3 of matrix $\boldsymbol{A}'$, and thus the transposed matrix is obtained.

The second thing to be learned about matrix multiplication is that matrices must be *conformable* to one another in order to multiply them. Matrix algebra is not like scalar algebra where any two

1 It is customary that the natural form of a vector is column form, and its row vector form is the transposed form. For example, the vector labelled as $\boldsymbol{f}$ above is the 3×1 column vector, and the vector labelled as $\boldsymbol{f}'$ is the 1×3 row vector.

numbers can be multiplied. Not every possible pair of matrices can be multiplied. This property of conformability is tersely explained in a quote from an introductory chapter on matrix algebra:

> To be conformable, the number of *columns* in the pre-multiplying matrix must be equal to the number of *rows* in the post-multiplying matrix (Brown et al., 2012, p. 91).

Suppose we wish to multiply matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ given above, *in that order*, $\boldsymbol{B}$ times $\boldsymbol{C}$. The number of columns in $\boldsymbol{B}$, the pre-multiplying matrix, must be equal to the number of rows in $\boldsymbol{C}$, the post-multiplying matrix. A quick way to check for conformability is to write the dimensions of the two matrices, the 2×4 matrix $\boldsymbol{B}$ and the 4×6 matrix $\boldsymbol{C}$, adjacent to one another, that is, $(2 \times 4)(4 \times 6)$. We see that the two "inner" dimensions (the number of columns in the first matrix and the number of rows in the second matrix) agree, indicating that the matrices can be multiplied. The two "outer" dimensions, 2 and 6 give the size of the resultant product matrix, that is, a 2×6 matrix will result.

Notice that vector $\boldsymbol{f}'$ and vector $\boldsymbol{d}$ are also conformable $(1 \times 3)(3 \times 1)$. The "inner" dimensions agree, and the "outer" dimensions indicate that the result will be a 1×1 matrix, that is, a scalar quantity. These two vectors will be used to demonstrate that the matrix multiplication process is essentially a method for obtaining a "sum of products" of the two adjacent vectors. To multiply two vectors, we take the row from the first, lay it side-by-side with the column of the second, multiply the corresponding elements and sum them up, as shown in Table A3.1 and as also shown below in matrix multiplication form.

$$\boldsymbol{f}'\boldsymbol{d} = \begin{bmatrix} 5 & 2 & 7 \end{bmatrix} \times \begin{bmatrix} 3 \\ 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 5 \times 3 + 2 \times 1 + 7 \times 4 \end{bmatrix} = 45$$

It is now clear why the number of columns in the pre-multiplying matrix must be equal to the number of rows in the post-multiplying matrix. The process consists of obtaining products of pairs of values and then summing them up—a Bravaisian product sum. *If the number of elements in the pre-multiplying vector were not equal to the number of elements in the post-multiplying vector, there would not be a complete set of pairs.*

This process is simple for vectors, but somewhat more complex when multiplying two matrices, in which there are multiple rows in the pre-multiplying matrix and multiple columns in the post-multiplying matrix. This more complex case of matrix multiplication requires obtaining the sum of products for each row of the pre-multiplying matrix paired with each column of the post-multiplying matrix. Consider, for example, the multiplication of matrix $\boldsymbol{A}$ and matrix $\boldsymbol{B}$. The dimensions of $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively, are 3×2 and 2×4, so the dimensions of the resultant matrix, which we call $\boldsymbol{D}$, will be 3×4. The two entries in row 1 of matrix **A** are paired with the two entries of column 1 of matrix **B**, and we obtain the sum of their products which is then entered into row 1 and column 1 of the product matrix $\boldsymbol{D}$ $(3 \times 5 + 2 \times 3 = 21)$. This process is followed for all

Table A3.1 The sum of products of vector $\boldsymbol{f}'$ and vector $\boldsymbol{d}$.

$\mathbf{5 \times 3 = 15}$
$\mathbf{2 \times 1 = 2}$
$\mathbf{7 \times 4 = 28}$
sum = **45**

combinations of the three rows of matrix $\boldsymbol{A}$ and the four columns of matrix $\boldsymbol{B}$, forming the resulting 3×4 product matrix $\boldsymbol{D}$.

$$\boldsymbol{D} = \boldsymbol{AB} = \begin{bmatrix} 3 & 2 \\ 1 & 4 \\ 6 & 2 \end{bmatrix} \times \begin{bmatrix} 5 & 1 & 7 & 2 \\ 3 & 3 & 5 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 3\times5+2\times3 & 3\times1+2\times3 & 3\times7+2\times5 & 3\times2+2\times4 \\ 1\times5+4\times3 & 1\times1+4\times3 & 1\times7+4\times5 & 1\times2+4\times4 \\ 6\times5+2\times3 & 6\times1+2\times3 & 6\times7+2\times5 & 6\times2+2\times4 \end{bmatrix}$$

$$= \begin{bmatrix} 21 & 9 & 31 & 14 \\ 17 & 13 & 27 & 18 \\ 36 & 12 & 52 & 20 \end{bmatrix}$$

These are the steps of matrix multiplication. Obviously, it can be a fairly tedious process to do by hand. One would hardly call it a labor-saving device. But tedious processes are exactly what computers are made for. When you let a spreadsheet do the work, matrix multiplication *is* a major labor-saving device as shall be demonstrated. Fortunately, Excel and most other spreadsheets can be used for matrix multiplication, making it readily accessible.[2]

One of the great advantages of doing a statistical analysis using matrix algebra, particularly when first learning a particular statistical method, is that one must know each of the logical and mathematical steps clearly in order to implement them in a matrix process. On the other hand, running a statistical analysis by the use of a statistical software package such as Stata or SAS is like a black box where the user often has only a rough idea of how the analysis is being carried out.

A3.2 Instructions for Doing Matrix Multiplication Using an Excel Spreadsheet

To get started using matrix operations on a spreadsheet, this section (A3.2) will provide step-by-step instructions in Excel for obtaining matrix $\boldsymbol{D}$ as the product of matrix $\boldsymbol{A}$ and matrix $\boldsymbol{B}$ as was shown mathematically above.

Step 1. Open an Excel spreadsheet and enter the $\boldsymbol{A}$ matrix in the location B2 to C4, and matrix $\boldsymbol{B}$ in the location E2:H3. The cells in Excel are identified by row and column using letters for columns and numbers for rows. For example, cell C4 of the spreadsheet is the cell in the third column and the fourth row. To enter matrix $\boldsymbol{A}$ into the B2:C4 location, place a 3 in cell B2, a 1 in cell B3, and a 6 in cell B4; then place a 2 in cell C2, a 4 in cell C3, and a 2 in cell C4. Enter the $\boldsymbol{B}$ matrix in the E2 to H3 location by a similar procedure. These two matrices could be labeled by entering "$\boldsymbol{A}$=" in cell A2, and "$\boldsymbol{B}$=" in cell D2.

2 There are many other matrix algebra processes, some of which are fairly complex and esoteric. Several additional matrix algebra procedures will be introduced in this book. Some of the most commonly used ones for data analysis and psychometrics are also explained in a form accessible to non-mathematicians in chapter 3 of the Brown et al. (2012) "simplest case" multivariate analysis book.

Step 2. The resulting product matrix $\boldsymbol{D}$ is a 3×4 matrix. A place could be reserved for matrix $\boldsymbol{D}$ in the location B6:E8, directly below matrix $\boldsymbol{A}$. To do this, highlight the 12 cells needed for matrix $\boldsymbol{D}$ by dragging your cursor from cell B6 to cell E8.

Step 3. While leaving the matrix $\boldsymbol{D}$ area highlighted, click on the "Formula Bar" at the top of the spreadsheet and type the following Excel command (without the quotes) into the Formula Bar: "=mmult (B2:C4,E2:H3)." The Excel keyword "**mmult**" stands for "matrix multiply."

Step 4. Still leaving the matrix $\boldsymbol{D}$ area highlighted, execute this typed command by holding down the "control" key and the "shift" key while you press the "enter" key.[3] If all of this goes well, the twelve values of the $\boldsymbol{D}$ product matrix will appear in cells B6:E8. The label "$\boldsymbol{D}$=" could now be entered into cell A6.

$$\boldsymbol{D} = \begin{bmatrix} 21 & 9 & 31 & 14 \\ 17 & 13 & 27 & 18 \\ 36 & 12 & 52 & 20 \end{bmatrix}$$

With that brief introduction to the process of multiplying matrices, the remaining sections of this appendix are devoted to illustrating how matrix multiplication can be used to create and analyze several descriptive statistics, including variance/covariance matrices.

A3.3 Calculating Variance from Raw Data by Three Matrix Methods

Shown in Table A3.2 is a matrix $\boldsymbol{X}$ containing a small dataset of scores of six persons on three tests—test A, test B, and test C. In preparation for demonstrating the calculation in Excel of a 3×3 variance/covariance matrix for these three tests in the next section (A3.4), the calculation of the variance of just one test, test A, will be demonstrated using three particular matrix multiplication methods. The transposed vector of raw scores on test A is $\boldsymbol{X}' = [8 \quad 6 \quad 5 \quad 5 \quad 4 \quad 2]$. The mean of these six values is 5. The transposed vector of deviation scores is therefore $\boldsymbol{x}' = [3 \quad 1 \quad 0 \quad 0 \quad -1 \quad -3]$, found by subtracting the mean of 5 from each of the six raw scores. The most basic matrix

Table A3.2 A 6 × 3 data matrix with column labels.

$$\boldsymbol{X} = \begin{array}{c} \begin{matrix} \text{Test A} & \text{Test B} & \text{Test C} \end{matrix} \\ \begin{bmatrix} 8 & 7 & 15 \\ 6 & 9 & 15 \\ 5 & 13 & 18 \\ 5 & 7 & 12 \\ 4 & 5 & 9 \\ 2 & 1 & 3 \end{bmatrix} \end{array}$$

3 This three-keystroke command is used for executing all Excel "array processing," such as matrix operations.

equation for calculating a variance can be stated simply as ***one Nth of the matrix product of $x'x$***, as demonstrated below.

$$S_A^2 = \left(\frac{1}{N}\right)\boldsymbol{x}'\boldsymbol{x} = \left(\frac{1}{6}\right)\begin{bmatrix}3 & 1 & 0 & 0 & -1 & -3\end{bmatrix}\begin{bmatrix}3\\1\\0\\0\\-1\\-3\end{bmatrix} = \left(\frac{1}{6}\right)(9+1+0+0+1+9) = \frac{20}{6} = 3.33$$

The resulting scalar[4] quantity of 3.33 is the sample variance. However, it is a biased sample variance, that is, it is a biased estimator of the corresponding population variance σ_A^2. In other words, in repeated sampling of variance calculations on samples of size $N = 6$, the expected value of the sample variances calculated by this process would systematically underestimate the true population variance by one Nth, that is, by one 6th. Given the very small sample size of $N = 6$ for our demonstration data, this estimate of 3.33 for the population variance has a large bias. The smaller the sample size, the larger the proportion of bias in estimating the population variance. To get an unbiased estimator of population variance, the product of the two vectors, $\boldsymbol{x}'\ \boldsymbol{x}$, should be multiplied by $1/(N-1)$ rather than $1/N$ as was shown in Appendix 2.

$$s_A^2 = \left(\frac{1}{N-1}\right)\boldsymbol{x}'\boldsymbol{x} = \left(\frac{1}{5}\right)\begin{bmatrix}3 & 1 & 0 & 0 & -1 & -3\end{bmatrix}\begin{bmatrix}3\\1\\0\\0\\-1\\-3\end{bmatrix} = \left(\frac{1}{5}\right)(9+1+0+0+1+9) = \frac{20}{5} = 4.0$$

In this appendix, and in Appendix 2, we use a lower-case s to designate the unbiased variance estimator s_A^2 in contrast to the capital S used to denote the biased estimator of the population variance S_A^2.

Besides the deviation score formula for unbiased variance just demonstrated, there is also a raw score formula for unbiased variance. It is slightly more complex, but not as susceptible to rounding error. Rather than multiplying deviation score vectors, it multiplies the raw score vectors $\boldsymbol{X}$ and $\boldsymbol{X}'$, and then "adjusts for" the mean by subtracting the product of the total (T) and the mean (M).

$$s_A^2 = \left(\frac{1}{N-1}\right)[\boldsymbol{X}'\boldsymbol{X} - TM'] = \left(\frac{1}{5}\right)\left[\begin{bmatrix}8 & 6 & 5 & 5 & 4 & 2\end{bmatrix}\begin{bmatrix}8\\6\\5\\5\\4\\2\end{bmatrix} - (30)(5)\right]$$

$$= \left(\frac{1}{5}\right)[(64+36+25+25+16+4) - 150] = \left(\frac{1}{5}\right)[170-150] = \frac{20}{5} = 4.0$$

4 Notice that the product of an $n \times 1$ column vector premultiplied by a $1 \times n$ vector is in fact a 1×1 matrix, that is, a single scalar quantity, referred to as a "dot product." On the other hand, notice that if you were to reverse this process and postmultiply the $n \times 1$ vector by the the $1 \times n$ vector, the product matrix would be $n \times n$, that is, a 6×6 matrix. This is referred to as the "matrix product" of these same two vectors.

There is also yet another matrix algebra formula for variance that uses both deviation scores and raw scores and is mid-way between the first two formulas in susceptibility to rounding error.

$$s_A^2 = \left(\frac{1}{N-1}\right)\boldsymbol{X}'\boldsymbol{x} = \left(\frac{1}{5}\right)\begin{bmatrix}8 & 6 & 5 & 5 & 4 & 2\end{bmatrix}\begin{bmatrix}3\\1\\0\\0\\-1\\-3\end{bmatrix}$$

$$= \left(\frac{1}{5}\right)(24 + 6 + 0 + 0 - 4 - 6) = \frac{20}{5} = 4.0$$

It may be surprising that this formula works, but it does.[5] The deviation score formula is susceptible to rounding error when using hand calculations for variance because when the mean is rounded to, for example, one place beyond the decimal point, each deviation score carries the error of that rounding, which is compounded by squaring. This formula still has some rounding errors but not as much because it does not square the errors.

A3.4 Calculating the 3 × 3 Covariance Matrix Using the Hybrid Formula

The 3×3 variance/covariance matrix, having variances in the diagonal locations (circled below) and covariances in the off-diagonal locations, may be calculated for the fictional dataset of three psychological tests using this hybrid raw-score/deviation-score formula. Notice that the symbol **S**, a bold capital "s" is used to signify the covariance matrix,[6] following the convention that matrices are indicated using bold upper-case letters.

$$\boldsymbol{S} = \begin{bmatrix} s_A^2 & s_{AB} & s_{AC} \\ s_{BA} & s_B^2 & s_{BC} \\ s_{CA} & s_{CB} & s_C^2 \end{bmatrix} = \left(\frac{1}{N-1}\right)\boldsymbol{X}'\boldsymbol{x} = \left(\frac{1}{5}\right)\begin{bmatrix} 8 & 6 & 5 & 5 & 4 & 2 \\ 7 & 9 & 13 & 7 & 5 & 1 \\ 15 & 15 & 18 & 12 & 9 & 3 \end{bmatrix} \cdot \begin{bmatrix} 3 & 0 & 3 \\ 1 & 2 & 3 \\ 0 & 6 & 6 \\ 0 & 0 & 0 \\ -1 & -2 & -3 \\ -3 & -6 & -9 \end{bmatrix}$$

5 The reason it works is because of the nature of a so-called centering matrix, that is, an $N \times N$ matrix $\boldsymbol{C_t}$ that will turn a raw data matrix $\boldsymbol{X}$ into a deviation score matrix $\boldsymbol{x}$ when the $\boldsymbol{X}$ matrix is pre-multiplied by it: $\boldsymbol{x} = \boldsymbol{C_t X}$. The centering matrix is symmetric and has the property of being idempotent, meaning that when it is matrix-multiplied by itself, the resulting product is equal to itself: $\boldsymbol{C_t C_t} = \boldsymbol{C_t}$. It follows then that one does not need to center both the pre-multiplying and post-multiplying $\boldsymbol{X}$ matrices, just one of them: $\boldsymbol{x}'\boldsymbol{x} = (\boldsymbol{C_t X})'(\boldsymbol{C_t X}) = \boldsymbol{X}'\boldsymbol{C}_t'\boldsymbol{C_t X} = \boldsymbol{X}'\boldsymbol{C_t X} = \boldsymbol{X}'\boldsymbol{x}$. The demonstration of this is found on pp. 103–108, and p. 155 of Brown et al. (2012).

6 We have sometimes referred to this as a "variance/covariance" matrix in this appendix since it contains variances in the diagonal elements and covariances in the off-diagonal elements. Paul Horst argues that this appears to be an unnecessarily clumsy designation. "It need only be remembered that the covariance of a variable with itself is a variance. In other words, a variance is a special case of a covariance where the two variables are identical. It is difficult to justify a complicated name when a simpler one will do" (Horst, 1968, p. 113).

$$= \left(\frac{1}{5}\right)\begin{bmatrix} 20 & 22 & 42 \\ 22 & 80 & 102 \\ 42 & 102 & 144 \end{bmatrix} = \begin{bmatrix} 4 & 4.4 & 8.4 \\ 4.4 & 16 & 20.4 \\ 8.4 & 20.4 & 28.8 \end{bmatrix}$$

We can check our calculations by copying and then pasting the "three tests" dataset into the data-entry spreadsheet of Stata and then executing the Stata command **correl atest btest ctest, covariance**. Stata will respond with the following covariance matrix output table which agrees with the results we obtained using matrix algebra in Excel.

```
. correl atest btest ctest, covariance
(obs=6)
```

	atest	btest	ctest
atest	4		
btest	4.4	16	
ctest	8.4	20.4	28.8

As indicated in Appendix 2, the Pearson product moment correlation coefficient (PPMCC) is defined as *standardized covariance*. We standardize a covariance by dividing it by the standard deviation of X and the standard deviation of Y:

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

The individual correlation coefficients of test A with test B, test A with test C, and test B with test C are calculated with this formula.

$$r_{ab} = \frac{S_{ab}}{S_a S_b} = \frac{4.4}{\sqrt{4}\sqrt{16}} = \frac{4.4}{(2)(4)} = \frac{4.4}{8} = 0.550$$

$$r_{ac} = \frac{S_{ac}}{S_a S_c} = \frac{8.4}{\sqrt{4}\sqrt{28.8}} = \frac{8.4}{(2)(5.37)} = \frac{8.4}{10.733} = 0.783$$

$$r_{bc} = \frac{S_{bc}}{S_b S_c} = \frac{20.4}{\sqrt{16}\sqrt{28.8}} = \frac{20.4}{(4)(5.37)} = \frac{20.4}{21.466} = 0.950$$

There is a matrix method for creating all coefficients of the correlation matrix $\boldsymbol{R}$ simultaneously. It is to both pre-multiply and post-multiply the covariance matrix by a *standardizing matrix* $\boldsymbol{E}$ according to the formula $\boldsymbol{R = ESE}$. The standardizing matrix $\boldsymbol{E}$ is a diagonal matrix, that is, a matrix with numerical entries in each of the diagonal positions, and zeros in the off-diagonal positions. The reciprocals of each of the three standard deviations become the entries for the three diagonal positions e_{11}, e_{22}, and e_{33} of the standardizing matrix $\boldsymbol{E}$.

$$e_{11} = s_a^{-1} = \frac{1}{s_a} = \frac{1}{2} = 0.500$$

$$e_{22} = s_b^{-1} = \frac{1}{s_b} = \frac{1}{4} = 0.250 \qquad \boldsymbol{E} = \begin{bmatrix} 0.500 & 0 & 0 \\ 0 & 0.250 & 0 \\ 0 & 0 & 0.186 \end{bmatrix}$$

$$e_{33} = s_c^{-1} = \frac{1}{s_c} = \frac{1}{5.37} = 0.186$$

Pre-multiplying the covariance matrix $\boldsymbol{S}$ by $\boldsymbol{E}$ standardizes it by rows, and post-multiplying it by $\boldsymbol{E}$ standardizes it by columns, thus creating the correlation matrix $\boldsymbol{R}$ by the matrix formula $\boldsymbol{R = ESE}$.

$$\boldsymbol{R = ESE} = \begin{bmatrix} 0.500 & 0 & 0 \\ 0 & 0.250 & 0 \\ 0 & 0 & 0.186 \end{bmatrix} . \begin{bmatrix} 4 & 4.4 & 8.4 \\ 4.4 & 16 & 20.4 \\ 8.4 & 20.4 & 28.8 \end{bmatrix} . \begin{bmatrix} 0.500 & 0 & 0 \\ 0 & 0.250 & 0 \\ 0 & 0 & 0.186 \end{bmatrix}$$

$$= \begin{bmatrix} 2.000 & 2.200 & 4.200 \\ 1.100 & 4.000 & 5.100 \\ 1.565 & 3.801 & 5.367 \end{bmatrix} . \begin{bmatrix} 0.500 & 0 & 0 \\ 0 & 0.250 & 0 \\ 0 & 0 & 0.186 \end{bmatrix} = \begin{bmatrix} 1.000 & 0.550 & 0.783 \\ 0.550 & 1.000 & 0.950 \\ 0.783 & 0.950 & 1.000 \end{bmatrix}$$

This matrix algebra result can be confirmed using the Stata command **correl atest btest ctest**.

```
. correl atest btest ctest
(obs=6)
```

	atest	btest	ctest
atest	1.0000		
btest	0.5500	1.0000	
ctest	0.7826	0.9503	1.0000

A3.5 Matrix Multiplication and the Equation for Probabilities-Weighting

At the beginning of Appendix 2, the "probabilities-weighting" method for calculating expected values, such as a mean or a variance, was demonstrated as a method for efficiently calculating such statistics from frequency-distribution data. A process was demonstrated for obtaining the mean as the sum of products of each of the unique score values multiplied by their respective probabilities. The example had four unique values, 4, 3, 2, and 1, with corresponding probabilities of 12/60, 21/60, 18/60, and 9/60, and the mean was found to be 2.6, as shown in Table A3.3.

$$mean = \overline{X} = E(X_i) = \sum p_i X_i = \left(\frac{12}{60}\right)4 + \left(\frac{21}{60}\right)3 + \left(\frac{18}{60}\right)2 + \left(\frac{9}{60}\right)1 = \frac{156}{60} = 2.6$$

This computational process is clear and simple in its matrix algebra form.

$$mean = \overline{X} = \boldsymbol{p}' \boldsymbol{X} = \begin{bmatrix} 12/60 & 21/60 & 18/60 & 9/60 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \\ 2 \\ 1 \end{bmatrix} = \frac{156}{60} = 2.6$$

The efficiency-of-computation argument is not quite as convincing with a small sample of data, but more compelling with a large sample when combined with a small set of unique values. Suppose, for example, a large pest control sales organization conducts a survey of work group size from 1000 random samples of their college student summer sales teams. The smallest work team consists of 38 persons and the largest consists of 172 persons. Table A3.4 shows the frequency distribution of these 1000 fictional sales teams and also the probability distribution values. Notwithstanding the large sample, there are only nine unique values for these 1000 teams. The mean team size for this sample can be found as the matrix product of the transposed vector of probabilities $\boldsymbol{p}'$ post-multiplied by the column vector of the nine unique values of team size, $\boldsymbol{X}$.

Note that the probabilities-weighting method can introduce rounding error if the probability values are not carried out to enough significant digits for the desired precision in the calculation of means or variances, with two or three more places beyond the decimal in the probabilities than the number of places in the result. In Table A3.4 example, the mean and the variance are calculated to three places, so the probabilities should be carried to five or six decimal places. However, since this example has a denominator of 1000 for the probabilities, the p_i values are at the exact precision level of the integer frequency values with three places for each decimal value.

Table A3.3 Frequency distribution of X scores.

X_i	f_i	p_i
4	12	12/60
3	21	21/60
2	18	18/60
1	9	9/60
$N = 60$		

Table A3.4 Frequency and probability distribution of sales team size.

X_i	f_i	p_i
4	38	0.038
5	62	0.062
6	87	0.087
7	131	0.131
8	172	0.171
9	171	0.171
10	153	0.153
11	108	0.108
12	78	0.078

$$mean = \overline{X} = E(X_i)$$

$$= \boldsymbol{p}'\boldsymbol{X} = \begin{bmatrix} 0.038 & 0.062 & 0.087 & 0.131 & 0.164 & 0.179 & 0.153 & 0.108 & 0.078 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \end{bmatrix}$$

$$= 8.47$$

Using this same process, the variance of team sizes across the 1000 sales teams can be calculated as the expected value of the squared deviation scores. The nine squared deviation scores (x_i^2) are first found by subtracting the mean from each of the team size values and then squaring the deviations.

$$x_1^2 = (4-8.47)^2 = (-4.47)^2 = 19.98 \quad x_2^2 = (5-8.47)^2 = (-3.47)^2 = 12.04$$
$$x_3^2 = (6-8.47)^2 = (-2.47)^2 = 6.10 \quad x_4^2 = (7-8.47)^2 = (-1.47)^2 = 2.16$$
$$x_5^2 = (8-8.47)^2 = (-0.47)^2 = 0.22 \quad x_6^2 = (9-8.47)^2 = (0.53)^2 = 0.28$$
$$x_7^2 = (10-8.47)^2 = (1.53)^2 = 2.34 \quad x_8^2 = (11-8.47)^2 = (2.53)^2 = 6.40$$
$$x_9^2 = (12-8.47)^2 = (3.53)^2 = 12.46$$

These deviation score values are entered into the expected value formula as the matrix product of the transposed vector of probabilities $\boldsymbol{p}'$ post-multiplied by the column vector of the nine squared deviation scores, $\boldsymbol{x^2}$.

$$S^2 = E(x_i^2) = \boldsymbol{p}'\boldsymbol{x^2}$$

$$= \begin{bmatrix} 0.038 & 0.062 & 0.087 & 0.131 & 0.172 & 0.171 & 0.153 & 0.108 & 0.078 \end{bmatrix} \begin{bmatrix} 19.98 \\ 12.04 \\ 6.10 \\ 2.16 \\ 0.22 \\ 0.28 \\ 2.34 \\ 6.40 \\ 12.46 \end{bmatrix} = 4.427$$

This variance created by the expected values formula is biased, but only slightly so with such a large N. The biased variance S^2 is multiplied by the conversion ratio to make it unbiased:

$$\hat{\sigma}^2 = \left(\frac{N}{N-1}\right)S^2 = \left(\frac{1000}{999}\right)4.427 = 4.432$$

The standard deviation is $\hat{\sigma} = \sqrt{4.432} = 2.105$. The histogram in Figure A3.1 shows the frequency distribution of each work group size with the mean work group size as the balance point of the

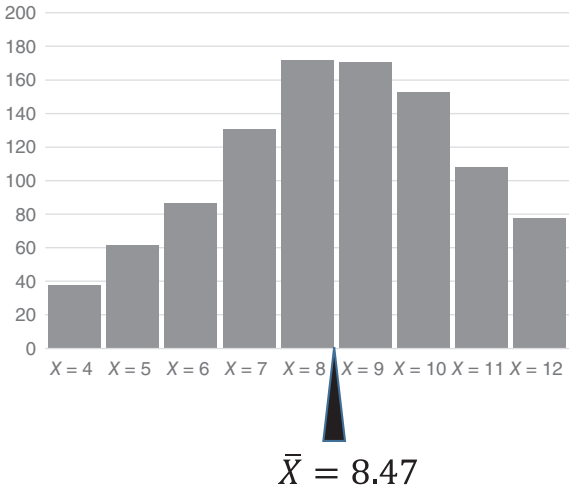


Figure A3.1 Histogram of the frequency distribution of work group sizes for nine work groups with the mean group size as the balancing point.

distribution. The upper boundary of workgroup size 12 on the right is 12.5, about two standard deviation units above the mean of 8.47. The lower boundary of the workgroup size of 4 on the left is 3.5, which has a standard score value of 2.364, which is a little larger than two standard deviation units below the mean, fitting the pattern of approximately 95% of the scores between +2 and −2 standard deviations from the mean.

$$Z = \frac{X_i - \overline{X}}{\hat{\sigma}} = \frac{3.5 - 8.47}{2.105} = \frac{-4.97}{2.105} = 2.361$$

A3.6 Summary

In this appendix, we have introduced the process of matrix multiplication and illustrated how it can be used to obtain covariance matrices and correlation matrices. In particular, we have shown how it could be used to obtain these matrices for a dataset containing several psychological tests. We have also demonstrated how to use matrix multiplication in combination with the "algebra of expectations" formula to efficiently calculate the mean and the variance of a large sample when there are a small number of unique score values.

A3.7 Study Questions for Appendix 3

Matrix Multiplication

Essay Questions

1 What do we mean when we say that variance is a *special case* of covariance (middle of the second paragraph of the appendix)? Can a variance be calculated using the covariance formula? Can a covariance be calculated using the variance formula? See also the quote from Paul Horst in footnote 6 of page 6 of this appendix.

2 What is the convention for specifying the dimensions of a matrix (paragraph 1, section A3.1)? Which is specified first, number of rows or number of columns? What is a "4 by 3" matrix, that is, a matrix which is specified as (4×3)?

3 What is a "scalar" (paragraph 1, section A3.1)? What is a vector? How does a vector differ from a scalar? How does a matrix differ from a vector?

4 How do we create the transpose of a matrix or vector (paragraph 3, section A3.1)? Suppose you transpose a 3×4 matrix. What will be the dimensions of the new matrix?

5 How can you determine whether two matrices are conformable for multiplication, that is, whether it is possible for the two matrices to be multiplied (paragraphs 5 and 6, section A3.1)? Once you have multiplied two conformable matrices, what will be the dimensions of the product of the two matrices? How many rows will it have? How many columns will it have? Suppose, for example, you have a 2×3 matrix $\boldsymbol{A}$, and a 3×5 matrix $\boldsymbol{B}$. Are they conformable for multiplication? What will be the dimensions of the product of these two matrices?

6 Explain why two matrices are not "conformable" for multiplication, that is cannot be multiplied, unless the number of columns in the pre-multiplying matrix is equal to the number of rows in the post-multiplying matrix (paragraphs 6 to 12 of section A3.1).

7 Why will the product matrix created by pre-multiplying a matrix by its own transpose always be symmetrical, with the upper triangular portion of the matrix (above the diagonal) being equivalent to the lower triangular portion of the matrix (below the diagonal), as it was shown to be in calculating the covariance matrix in paragraph 1 of section A3.4?

8 Explain how any vector pre-multiplied or post-multiplied by its own transpose could either produce a "dot product" or a "matrix product" as discussed in footnote 4.

9 Explain the simple process of pre-multiplying a column vector by its own transpose.

Calculation Questions

For each of the following questions, you may do the calculations either with a pocket calculator or using the matrix capabilities of Excel.

1 Pre-multiply the following column vector $\boldsymbol{a}$ by its transpose to obtain a sum of squares, $SS = \boldsymbol{a}'\boldsymbol{a}$.

$$\boldsymbol{a} = \begin{bmatrix} 9 \\ 10 \\ 7 \\ 12 \\ 8 \end{bmatrix}$$

2 Post-multiply matrix $\boldsymbol{A}$ by matrix $\boldsymbol{B}$ to obtain matrix $\boldsymbol{C}$.

$$\boldsymbol{A} = \begin{bmatrix} 1 & 3 & 2 \\ 5 & 7 & 2 \end{bmatrix} \quad \boldsymbol{B} = \begin{bmatrix} 3 & 4 & 3 & 2 \\ 1 & 1 & 4 & 2 \\ 5 & 6 & 2 & 5 \end{bmatrix}$$

3 For the small dataset of measures of six persons on three psychological tests (Test A, Test B, and Test C) found in Table A3.2 of section A3.3 use the deviation-score-based (definitional) formula to obtain the covariance matrix.

$$\boldsymbol{S} = \left(\frac{1}{N-1}\right)\boldsymbol{x}'\boldsymbol{x}$$

4 Calculate the correlation matrix for this same small dataset from Table A3.2 of section A3.3, using the matrix formula $\boldsymbol{R = ESE}$, given in paragraph 6 of section A3.4.

5 Suppose that you have a slightly larger dataset of measures of seven persons on three additional psychological tests, Test D, Test E, and Test F, as shown in the table below. Calculate the covariance matrix and the correlation matrix for this dataset and then verify your results using Stata.

Test D	Test E	Test F
12	6	18
9	3	12
7	4	11
7	7	14
7	11	18
4	13	17
3	12	15

6 Your research team creates a psychological test consisting of five binary items, where the score on the test is simply the number of items answered correctly. The test is administered to 160 students at the university producing the results shown in the table below, with the X_i column being the values of the scores themselves, f_i being the frequency of each score, and p_i being the probability of each score (frequency/sum). Use the method shown in paragraphs 1 through 3 of section A3.5, the matrix algebra version of the "probabilities-weighting" expected values equation, to efficiently calculate the mean of the scores. Use Excel to also create a histogram of the scores.

X_i	f_i	p_i
5	27	0.16875
4	59	0.36875
3	46	0.28750
2	13	0.08125
1	9	0.05625
0	6	0.03750
Sums	160	1.00000

7 Calculate the biased variance of these 160 scores by first obtaining the squared deviation of each of these six scores from the mean calculated in question 6. The biased sample variance can now be obtained by pre-multiplying the column vector of these squared deviation scores $\boldsymbol{x}^2$, by the row vector of probabilities $\boldsymbol{p}$, according to the formula $S^2 = \boldsymbol{p}'\boldsymbol{x}^2$. Finally, use the conversion ratio formula to convert this biased variance into an unbiased estimate of the population variance. These calculations could be accomplished using a hand calculator or with a spreadsheet such as Excel.

Appendix 4

Bivariate Regression Analysis

APPENDIX MENU

A4.1 Scatterplots and Regression Lines Comparing Three Datasets, One with High, One with Medium, and One with Low Predictive Strength, 520
A4.2 Computational Steps Underlying Bivariate Regression, 521
- A4.2.1 Step 1. Calculate the Covariance Matrix, the Standard Deviation, and the Correlation Coefficient, 521
- A4.2.2 Step 2. Calculate the Slope and the *Y* Intercept, and Write the Predictive Equation, 523
- A4.2.3 Step 3. Use the Predictive Equation to Create Predicted Scores on *Y*, 524
- A4.2.4 Step 4. Calculate SSR, the Sum of Squares Regression, 526
- A4.2.5 Step 5. Calculate the Errors of Prediction by Subtracting Predicted Scores $\hat{Y}$ from Actual *Y* Scores, 526
- A4.2.6 Step 6. Calculate SSE, the Sum of Squares Error, 527
- A4.2.7 Step 7. Confirm that $SST = SSR + SSE$, and also that $r^2 = SSR/SST$, and that $1 - r^2 = SSE/SST$, 527
- A4.2.8 Step 8. Calculate the ANOVA Table Statistics: The Three Mean Squares, the *F*-ratio, and the Standard Error of the Estimate, $s_{y.x}$, 528

A4.3 Calculating Bivariate Regression Using Stata, 529
A4.4 Study Questions for Appendix 4, 531

Bivariate regression, the topic of this appendix, is the simplest case of least squares regression. It is called "bivariate" because there are only two variables, a single predictor variable, and a single criterion variable. In sections §28 and §30 of Chapter 3 and in Appendix 5, a more complex and powerful variety of least squares regression is introduced—*multiple regression*. In multiple regression, a single *Y* variable is predicted from two or more predictor variables in combination with one another. Learning the principles of bivariate regression analysis in this section will facilitate learning multiple regression in Appendix 5 and in Chapter 3.

The purpose of bivariate linear regression analysis is to predict values on one variable, *Y*, from values on another variable, *X*. The Pearson product moment correlation coefficient (PPMCC) provides the basis upon which this can be done. To the extent that the correlation coefficient between the two variables is strong, one can make reasonably accurate predictions from one variable to the other. The *Y* variable in bivariate linear regression is the one that is predicted. It is often referred to as the *criterion* variable. The *X* variable is the *predictor* variable. When referring to the process, we say that we are *regressing variable Y* (the criterion variable) *onto variable X* (the predictor variable).

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

This is accomplished through the use of a *regression line*, which is the modern counterpart of Galton's 'reversion line' discussed in the third to last paragraph of subsection 1.1.4.2 of Chapter 1.

A4.1 Scatterplots and Regression Lines Comparing Three Datasets, One with High, One with Medium, and One with Low Predictive Strength

Shown in Figure A4.1 are three bivariate scatterplots, one for each of three individual bivariate regression analyses. A predictive regression line is also shown for each. Scatterplot C in the rightmost panel has a reasonably strong linear trend. That is to say, the six bivariate observations for Dataset C are quite closely focused along a straight line in an X/Y plot, such that the values of variable Y can be predicted well from the values in variable X. In Dataset C, those persons with higher scores on X also tend to have higher scores on Y. Scatterplot B, on the other hand, has a medium level of predictive strength. Scatterplot A, in the leftmost panel, is a dataset with a low level of predictive strength, and weak linear trend.

The numerical X and Y values for Datasets A–C that were used to create these scatterplots and regression lines are shown in Figure A4.1, with six persons (labeled A–F) in each dataset. Each of the six persons has both an X score and also a Y score. For pedagogical reasons, we have constructed these three datasets to be readily comparable to one another. That is, all three datasets have exactly the same numerical values on variable X. All three values are also in the same order from left to right across the observations, from observation F to observation A (F = 8, E = 6, D = 5, C = 5, B = 4, and A = 2). The three datasets also have exactly the same numerical values on variable Y

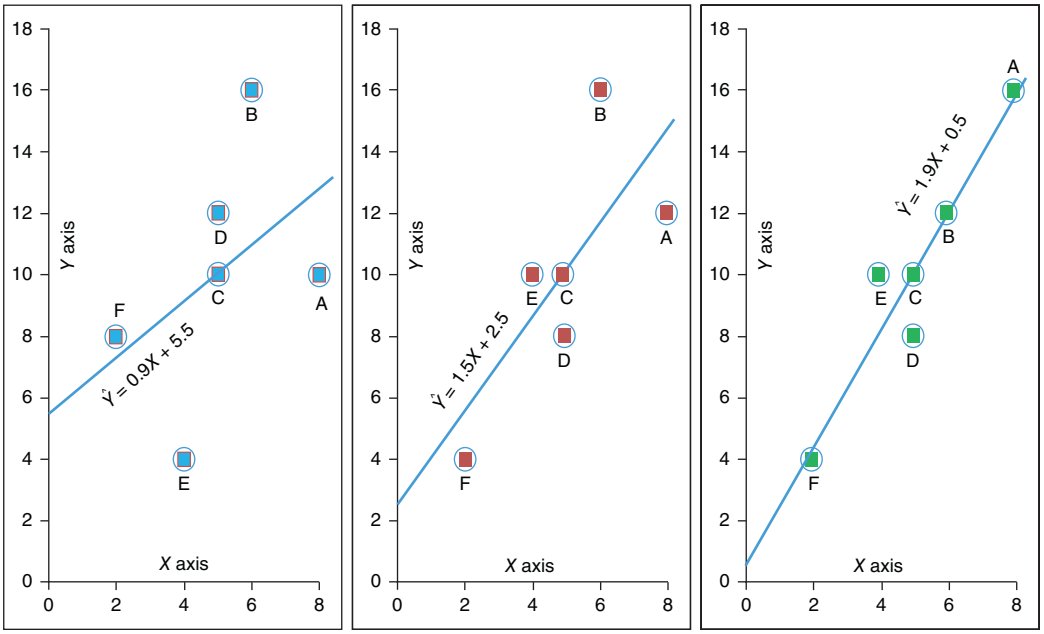


Figure A4.1 Three bivariate scatterplots illustrating three contrasting levels of predictive strength. The regression line for each dataset is also shown, superimposed upon its scatterplot.

but with different numerical-value locations across the three datasets. Dataset C has the order (A = 16, B = 12, C = 10, E = 10, D = 8, and F = 4) from highest to lowest for the six values. Dataset B has the order (B = 16, A = 12, C = 10, E = 10, D = 8, and F = 4) for a different pattern of these same six numerical values; and Dataset A has yet another order (B = 16, D = 12, C = 10, A = 10, F = 8, and E = 4) for these same six numerical values. The only thing that differs among the three datasets is *the way in which the X and Y scores are paired*, that is, their covariance—their index of relatedness. In Dataset C, the *X*/*Y* bivariate pairs have a strong relationship—a strong linear trend—and therefore a large covariance. Those persons who have high scores on *X*, for example, person A and person B, also tend to have high scores on *Y*. Similarly, those persons who are low on *X* also tend to be low on *Y*. On Dataset B, the linear trend is not quite so strong. The scores on *Y* cannot be predicted as well from *X* for Dataset B. And, on Dataset A, the linear trend is even weaker. The computations underlying bivariate regression analysis will now be demonstrated in Section A4.2 for these three datasets that differ in predictive strength.

A4.2 Computational Steps Underlying Bivariate Regression

A4.2.1 Step 1. Calculate the Covariance Matrix, the Standard Deviation, and the Correlation Coefficient

The deviation score matrix algebra formula will be used to calculate the covariance matrix, first on Dataset C. The sum of scores for variable *X* is 30. So, the mean is 5, and the deviation scores on variable X are 3, 1, 0, 0, −1, −3, as the reader can verify by subtracting 5 from each of the six X scores, (which from right to left are 8, 6, 5, 5, 4, 2 as shown in the graph for Dataset C in Figure A4.1). Deviation scores are "centered" scores, that is, they are centered upon their mean, and therefore their sum is zero. In other words, the sum of the negative-valued deviation scores is equal to the sum of the positive-valued ones. The sum of variable *Y* scores is 60, the mean is 10, and the deviation scores are 6, 2, 0, −2, 0, −6, for persons A to F, respectively. Using the matrix algebra deviation score formula from Appendix 3, Sections A3.3 and A3.4, we calculate the ***CSSCP*** matrix (which stands for the "centered sum of squares and cross products" matrix) for these two variables as found in Dataset C.

$$\boldsymbol{CSSCP} = \boldsymbol{x}'\boldsymbol{x} = \begin{bmatrix} 3 & 1 & 0 & 0 & -1 & -3 \\ 6 & 2 & 0 & -2 & 0 & -6 \end{bmatrix} \begin{bmatrix} 3 & 6 \\ 1 & 2 \\ 0 & 0 \\ 0 & -2 \\ -1 & 0 \\ -3 & -6 \end{bmatrix} = \begin{bmatrix} 20 & 38 \\ 38 & 80 \end{bmatrix} = \begin{bmatrix} SS_x & CP_{xy} \\ CP_{yx} & SS_y \end{bmatrix}$$

Each centered sum of squares (*CSS*) value and centered cross product (*CCP*) value is divided by the degrees of freedom (*df*) to obtain the variance/covariance matrix **S**. This is accomplished by multiplying each by 1/*df*. To clarify, variance (the mean square) is found by dividing variation[1]

1 The term 'variation' is often used in statistical practice to refer to *CSS*, the sum of squares of the centered (deviation) scores. These centered scores are the scores that appear in the ***CSSCP*** equation above, both in the row-oriented premultiplying matrix ***x'***, and also in the column-oriented postmultiplying matrix ***x***. Notice that the first column of matrix ***x***, the postmultiplying matrix, is the deviation scores from variable *X* and the second column is the deviation scores from variable *Y*, as they appear in previous pages.

(the CSS) by $N-1$ degrees of freedom. Similarly, the covariance (the mean product) is found by dividing the CCP, or sum of products,[2] by the degrees of freedom.

$$\boldsymbol{S} = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{yx} & s_y^2 \end{bmatrix} = \left(\frac{1}{df}\right)\boldsymbol{CSSCP} = \left(\frac{1}{N-1}\right)\boldsymbol{x}'\boldsymbol{x} = \left(\frac{1}{5}\right)\begin{bmatrix} 20 & 38 \\ 38 & 80 \end{bmatrix} = \begin{bmatrix} 4 & 7.6 \\ 7.6 & 16 \end{bmatrix}$$

Since the variance for X, the first variable, is 4, its standard deviation is 2. Likewise, since the variance of Y is 16, its standard deviation is 4. The Pearson Product Moment Correlation Coefficient (PPMCC) is defined mathematically as the covariance divided by its two standard deviations.

$$r = \frac{s_{xy}}{s_x s_y} = \frac{7.6}{(2)(4)} = \frac{7.6}{8} = 0.95$$

The correlation coefficient for Dataset C is fairly strong, substantially stronger than those for Datasets A and B, which will now be calculated for comparison. In comparing the three datasets, notice that the X scores are exactly the same for all three—the same values and in the same ordinal arrangement. The X deviation scores for both datasets, A and B, are 3, 1, 0, 0, −1, −3, just as they were for Dataset C. However, for Dataset B, the Y deviation scores are 2, 6, 0, −2, 0, −6. The values are the same as the Y values in Dataset C, but the order is not. The covariance is therefore calculated to be 6 for Dataset B, somewhat smaller than the covariance of 7.6 for Dataset C.

$$s_{xy} = \left(\frac{1}{N-1}\right)\boldsymbol{x}'\boldsymbol{y} = \left(\frac{1}{6-1}\right)\begin{bmatrix} 3 & 1 & 0 & 0 & -1 & -3 \end{bmatrix}\begin{bmatrix} 2 \\ 6 \\ 0 \\ -2 \\ 0 \\ -6 \end{bmatrix} = \left(\frac{1}{5}\right)(30) = 6$$

The PPMCC for Dataset B is calculated to be 0.75.

$$r = \frac{s_{xy}}{s_x s_y} = \frac{6}{(2)(4)} = \frac{6}{8} = 0.75$$

The Y deviation scores for Dataset A are found to be 0, 6, 0, 2, −6, −2. These are the same as the Y deviation score values for Dataset B, but organized in a different order, and therefore the covariance will differ. The covariance for Dataset A is calculated to be 3.6.

$$s_{xy} = \left(\frac{1}{N-1}\right)\boldsymbol{x}'\boldsymbol{y} = \left(\frac{1}{6-1}\right)\begin{bmatrix} 3 & 1 & 0 & 0 & -1 & -3 \end{bmatrix}\begin{bmatrix} 0 \\ 6 \\ 0 \\ 2 \\ -6 \\ -2 \end{bmatrix} = \left(\frac{1}{5}\right)(18) = 3.6$$

The correlation coefficient for Dataset A is therefore found to be 0.45.

$$r = \frac{s_{xy}}{s_x s_y} = \frac{3.6}{(2)(4)} = \frac{3.6}{8} = 0.45$$

2 "Cross product" is the 20th and 21st centuries' term for the "product sum" demonstrated by Bravais in his 1846 paper, as discussed the third to last paragraph of subsection 1.1.4.2 of Chapter 1. It is also sometimes referred to as the sum of products.

Squaring the correlation coefficient creates a *coefficient of determination*, also called an r^2 value, which indicates what proportion of the variance in variable Y can be predicted by regressing it onto variable X. For Dataset C, with a correlation coefficient of 0.95, we account for a high percentage of variance, $cd = r^2 = 0.9025$. In other words, 90¼% of the variance in Y is accounted for by regressing it onto X. On the other hand, Dataset B has an r^2 value of $0.75^2 = 0.5625$, accounting for only a little over half of the variance; and in Dataset A, the bivariate regression only accounts for 20¼% of the variance, with $cd = r^2 = 0.45^2 = 0.2025$.

If we subtract the coefficient of determination from 1.00, we obtain the coefficient of nondetermination. For example, we calculate the coefficient of nondetermination for Dataset C.

$$cnd = 1 - r^2 = 1 - 0.9025 = 0.0975$$

The coefficient of nondetermination (*cnd*) is interpreted as the proportion of variance *not* accounted for by the regression equation. The square root of the coefficient of nondetermination is the coefficient of alienation, which is the ratio of the standard error of the estimate (that is, the standard deviation of the errors away from the regression line) to the standard deviation of Y.

$$coefficient\_of\_alienation = \sqrt{1 - r^2} = \sqrt{0.0975} = 0.3122$$

These four quantities *r*, *cd*, *cnd*, and *ca*, can be collected into a small table that we refer to as a "correlational fourfold." It consists of four quantities which are used together to explain the interpretation of each PPMCC.

PPMCC Pearson's *r*	Coefficient of determination	Coefficient of nondetermination	Coefficient of alienation
R	r^2	$1 - r^2$	$\sqrt{1 - r^2}$

This is the fourfold table for Dataset C.

PPMCC Pearson's *r*	Coefficient of determination	Coefficient of nondetermination	Coefficient of alienation
0.9500	0.9025	0.0975	0.3122

Having obtained the means and standard deviations of X and Y, and also the correlation coefficients between X and Y for each of the three datasets, it is now possible to calculate the regression parameters b and a. These can be used to construct the regression equations for each dataset to compare to the regression lines and their equations shown in Figure A4.1. The calculations will be demonstrated for only Dataset C, and then left as an exercise for the student to complete for the other two datasets (in *Calculation Questions* 1 and 2 in the *Study Questions* at the end of this appendix).

A4.2.2 Step 2. Calculate the Slope and the *Y* Intercept, and Write the Predictive Equation

The equation for a bivariate regression line is of the form $\hat{Y} = bX + a$, where b is the slope, a is the "Y intercept," and $\hat{Y}$ is the predicted score for Y. The slope is obtained as the product of the correlation coefficient and the ratio of the standard deviations, with the standard deviation of Y being divided by the standard deviation of X. The formula is:

$$b = r\left(\frac{s_y}{s_x}\right) \tag{A4.1}$$

For Dataset C, the slope is calculated to be 1.90.

$$b = r\left(\frac{s_y}{s_x}\right) = (0.95)\left(\frac{4}{2}\right) = 1.90$$

The Y intercept is the location where the regression line "intercepts," or in other words, crosses, the Y axis. The formula for the Y intercept is the mean of Y minus the slope times the mean of X.

$$a = \overline{Y} - b\overline{X} \tag{A4.2}$$

For Dataset C, the intercept is calculated to be 0.5.

$$a = 10 - (1.90)(5) = 10 - 9.5 = 0.5.$$

The regression equation for Dataset C can now be written using these values for the slope b and the intercept a.

$$Y_i = 1.90\, X_i + 0.5 \tag{A4.3}$$

Equation (A4.3) is used to calculate the regression line for Dataset C in the following subsection. The symbol $\hat{Y}$ ("Y-hat") stands for the predicted scores in contrast to the actual Y scores. The "hat" over the symbol for a variable indicates that the variable is in some sense statistically optimal. In this case, it indicates that these particular bivariate regression-predicted scores are optimal in a "least squares" sense as explained in footnote 3.

A4.2.3 Step 3. Use the Predictive Equation to Create Predicted Scores on *Y*

Each of the six predicted scores on Y is calculated by entering its X value into this regression equation and solving for $\hat{Y}_i$,

$$\hat{Y}_1 = 1.9\,(X_1) + 0.5 = 1.9\,(8) + 0.5 = 15.2 + 0.5 = 15.7$$

$$\hat{Y}_2 = 1.9\,(X_2) + 0.5 = 1.9\,(6) + 0.5 = 11.4 + 0.5 = 11.9$$

$$\hat{Y}_3 = 1.9\,(X_3) + 0.5 = 1.9\,(5) + 0.5 = 9.5 + 0.5 = 10.0$$

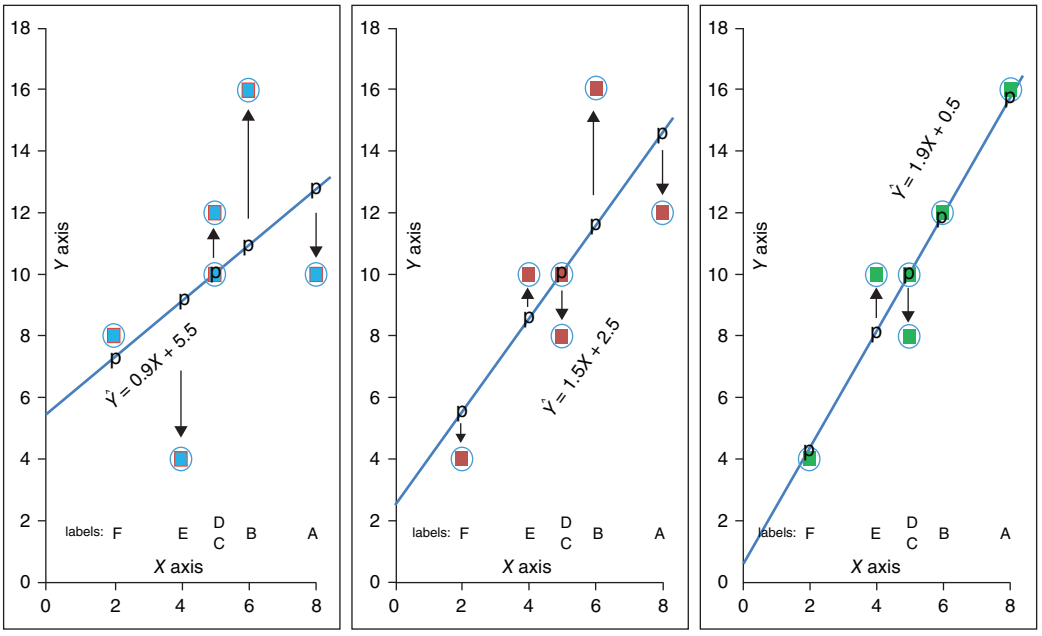
$$\hat{Y}_4 = 1.9\,(X_4) + 0.5 = 1.9\,(5) + 0.5 = 9.5 + 0.5 = 10.0$$

$$\hat{Y}_5 = 1.9\,(X_5) + 0.5 = 1.9\,(4) + 0.5 = 7.6 + 0.5 = 8.1$$

$$\hat{Y}_6 = 1.9\,(X_6) + 0.5 = 1.9\,(2) + 0.5 = 3.8 + 0.5 = 4.3$$

The bivariate scatterplots and the regression lines for all three datasets are again shown in Figure A4.2 (three-panel figure), but this time with three additions. The first addition is that we have placed a "**p**" (for "predicted value") at the location on the regression line where the predicted score for each of the six persons falls. Notice that each predicted score lies precisely on the regression line. The second addition is that we have placed arrows from the predicted scores to their corresponding observed scores to show the extent of deviation, upward or downward, between predicted score and actual score. The third addition is that we have placed grey person labels slightly above the X axis at the bottom of each scatterplot to indicate where the X score of each of the six persons is located. This is to facilitate identifying each deviation of a predicted score from a particular actual score.

We can now explain the construction and the meaning of the regression line for Dataset C. As an explanation of how to read the "**p**" symbols along the regression line to identify predicted scores,



Scatterplot for Dataset A with a regression line: low strength prediction

Scatterplot B for Dataset A with a regression line: medium strength prediction

Scatterplot C for Dataset A with a regression line: high strength prediction

Figure A4.2 The three bivariate scatterplots from Figure A4.1 are shown here but with a **p** on each regression line indicating the location of the predicted scores for each of the six persons. The vertical arrows above or below the marker for each person indicate the six error scores, that is, the difference between each predicted score and its corresponding actual score.

look at the scatterplot for Dataset C at the far right. The first bivariate predicted score (the one for Person A) is at the location of "$X = 8$ and $Y = 15.7$," on the extreme upper right of the regression line. The other five predicted score locations for Persons B through F are: B at "$X = 6$ and $Y = 11.9$," C at "5 and 10," D at "5 and 10," E at "4 and 8.1," and F at "2 and 4.3." The regression equation enables us to fit a straight line, a least squares[3] straight line, to the six empirical bivariate scores, and the predicted scores $\hat{Y}_i$ will all fall precisely on that line.

It can be shown that the slope of this regression line is indeed 1.9. For Dataset C, the difference between the highest and the lowest of the Y-predicted scores is $15.7 - 4.3 = 11.4$, and the difference between the highest and the lowest of the X scores is $8 - 2 = 6$. The "rise over run" ratio of these two is 11.4/6 which equals 1.9, the calculated slope of this line. Also, the regression line drawn through these six "**p**" bivariate predicted points can be seen to cross the Y axis at a value of 0.5 the Y intercept, a. We can numerically verify this intercept value more precisely by solving the predictive equation with zero entered for the value of X. That is to say that when the X value is set to zero, the corresponding $\hat{Y}$ value is 0.5, the Y intercept.

$$\hat{Y} = 1.9X + 0.05 = 1.9(0) + 0.05 = 0.5$$

3 The regression line is a least squares line in the sense that the sum of squared deviations of observed Y scores from the corresponding predicted $\hat{Y}$ scores on the regression line are less than from any other line. The student is invited to verify this in question 8 of the *Calculation Questions* at the end of this appendix.

A4.2.4 Step 4. Calculate SSR, the Sum of Squares Regression

The next step is to obtain the *sum of squares regression*, SSR, for these predicted scores. The mean of the $\hat{Y}$ predicted scores is the same as the mean of the observed Y scores, which was calculated to be 10. This is subtracted from each of the predicted scores to obtain deviation scores for Persons A through F, respectively, of 5.7, 1.9, 0.0, 0.0, −1.9, and −5.7. We will use the symbol $\hat{\mathbf{y}}$ to represent the deviation score form of the predicted scores $\hat{Y}$. Using matrix algebra to pre-multiply the column vector $\hat{\boldsymbol{y}}$ of these predicted scores by its own transpose, $\hat{\boldsymbol{y}}'$, creates *SSR*, the sum of squares regression.

$$SSR = \hat{\mathbf{y}}'\,\hat{\mathbf{y}} = \begin{bmatrix} 5.7 & 1.9 & 0 & 0 & -1.9 & -5.7 \end{bmatrix} \begin{bmatrix} 5.7 \\ 1.9 \\ 0 \\ 0 \\ -1.9 \\ -5.7 \end{bmatrix} = 72.2 \tag{A4.4}$$

This SSR value can be used to calculate the coefficient of determination, r^2, as the ratio of SSR, the sum of squares regression, to SST, the sum of squares total.

$$r^2 = \frac{SSR}{SST} = \frac{72.2}{80} = 0.9025 \tag{A4.5}$$

This helps clarify the meaning of a PPMCC. It is best interpreted by squaring it to create a coefficient of determination, which indicates how many variation units are predictable out of the total number of variation units in Y.

In a similar manner, all of these calculations could now be carried out on Datasets A and B. The slopes and intercepts can be calculated; the corresponding predicted scores can be calculated; the graphical renditions of slopes and intercepts of the regression lines can be verified; the *SSR* values (sum of squares regression), and the r^2 values can be calculated as they were here for Dataset C. In fact, that is your very task in questions C1 and C2 of the *Calculation Questions* at the end of this mathematical appendix, to calculate and verify each of these things for Datasets A and B.

A4.2.5 Step 5. Calculate the Errors of Prediction by Subtracting Predicted Scores $\hat{Y}$ from Actual Y Scores

The next step is to obtain the errors-of-prediction scores, which are symbolized by the Greek letter epsilon (ε). They are calculated by subtracting the $\hat{Y}_i$ predicted scores from the Y_i observed scores.

$$\varepsilon_i = Y_i - \hat{Y}_i \tag{A4.6}$$

The six errors-of-prediction scores for Dataset C are calculated using Eq. (A4.6).

$$\varepsilon_1 = Y_1 - \hat{Y}_1 = 16 - 15.7 = 0.3$$

$$\varepsilon_2 = Y_2 - \hat{Y}_2 = 12 - 11.9 = 0.1$$

$$\varepsilon_3 = Y_3 - \hat{Y}_3 = 10 - 10 = 0$$

$$\varepsilon_4 = Y_4 - \hat{Y}_4 = 8 - 10 = -2$$
$$\varepsilon_5 = Y_5 - \hat{Y}_5 = 10 - 8.1 = 1.9$$
$$\varepsilon_6 = Y_6 - \hat{Y}_6 = 4 - 4.3 = -0.3$$

A4.2.6 Step 6. Calculate SSE, the Sum of Squares Error

The errors-of-prediction scores can be used to calculate the *sum of squares error*, SSE. These six error scores are already in deviation score form, which can be verified by noticing that their values sum up to zero. This will always be true for errors of prediction. They always have a mean of zero. The matrix algebra formula for SSE is simple. SSE is just the column vector $\boldsymbol{\varepsilon}$ pre-multiplied by its transpose $\boldsymbol{\varepsilon}'$.

$$\boldsymbol{SSE} = \boldsymbol{\varepsilon}'\,\boldsymbol{\varepsilon} = \begin{bmatrix} 0.3 & 0.1 & 0.0 & -2.0 & 1.9 & -0.3 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.1 \\ 0.0 \\ -2.0 \\ 1.9 \\ -0.3 \end{bmatrix} = 7.8 \tag{A4.7}$$

A4.2.7 Step 7. Confirm that *SST* = *SSR* + *SSE*, and also that r^2 = *SSR*/*SST*, and that $1 - r^2$ = *SSE*/*SST*

In Appendix 1, it was taught that one of the main reasons that statistical science is based upon variance, a squared-units index of spread, is because, as noted by the Pythagoreans, squared units can be partitioned into parts. Nowhere is that more clearly illustrated than in bivariate regression. The total variation in Y scores, SST, can be partitioned into that part of the variation that is predictable, SSR, and that part of the variation that is not predictable, SSE, the error variation. The two parts are additive as expressed in Eq. (A4.8).

$$SST = SSR + SSE \tag{A4.8}$$

This additive decomposition of *SST* into the predicted and the unpredicted parts is well illustrated with the sum of squares from Dataset C, where *SST* is equal to 80, *SSR* is equal to 72.2, and *SSE* is equal to 7.8.

$$SST = 80 = 72.2 + 7.8 = SSR + SSE$$

In the same way that the coefficient of determination, r^2, can be calculated as the ratio of *SSR* to *SST* (as shown in Eq. (A4.5)), the coefficient of nondetermination, $1 - r^2$, can be calculated as the ratio of *SSE* to *SST* (as shown in Eq. (A4.9)).

$$1 - r^2 = \frac{SSE}{SST} = \frac{7.8}{80} = 0.0975 \tag{A4.9}$$

Clearly, the coefficient of determination (0.9025 for Dataset C) and the coefficient of nondetermination (0.0975) are additive and sum to 1.00, just as SSR and SSE are additive and sum to SST. This becomes particularly clear if we divide both sides of Eq. (A4.8) by SST.

$$r^2 + \left(1 - r^2\right) = \frac{SSR}{SST} + \frac{SSE}{SST} = \frac{SST}{SST} = 1$$

Or, to express it numerically:

$$0.9025 + 0.0975 = \frac{72.2}{80} + \frac{7.8}{80} = \frac{80}{80} = 1$$

A4.2.8 Step 8. Calculate the ANOVA Table Statistics: The Three Mean Squares, the *F*-ratio, and the Standard Error of the Estimate, $s_{y.x}$

In bivariate regression, it is actually the variation quantities (the sums of squares) rather than the variances that best illustrate the additivity of predicted and error partitions. However, to complete the presentation of bivariate regression, the next step is to convert the variations into variances, by dividing them by their degrees of freedom. The total degrees of freedom is given as $N - 1$, which is straightforward. There are N observations but only $N - 1$ of them are free to vary. The "last" one is constrained since the deviations must sum to zero. The equation for the total variance for this regression analysis, referred to as mean square total (MST), is found by dividing SST by its degrees of freedom, $N - 1$.

The degrees of freedom for the regression sum of squares, SSR, is $df_r = h - 1$, where h is the number of independent parameters that are used to create the regression line, which for the bivariate case is 2, the slope and the intercept. The degrees of freedom is $h - 1$ because the predicted scores, like the observed scores are constrained by their mean, that is, the sum of the predicted scores around their mean is zero. The degrees of freedom for the error sum of squares, *SSE*, is $df_e = N - h$ because the deviation of the N observed scores from the regression line is constrained by the two parameters used to create the regression line in bivariate regression.[4] Using these degrees of freedom for each of the three sums of squares yields the three mean square estimates for bivariate regression.

$$MSR = \frac{SSR}{h-1} = \frac{72.2}{1} = 72.2$$

$$MSE = \frac{SSE}{N-h} = \frac{7.8}{4} = 1.95$$

$$MST = \frac{SST}{N-1} = \frac{80}{5} = 16$$

An *F* ratio can be calculated from MSR and MSE to test the regression model for Dataset C for statistical significance.

$$F = \frac{SSR}{SSE} = \frac{72.2}{1.95} = 37.03$$

This F ratio has 1 degree of freedom in the numerator ($df_r = 1$), and 4 degrees of freedom in the denominator. So, the 0.01 level critical ratio is $F(0.01,1,4) = 21.20$. The obtained F ratio exceeds that value. So, the regression model is significant at the 0.01 level.

4 Perhaps one of the most helpful explanations of degrees of freedom in regression models is given in section 7.4 of Kirk (2013), "Estimating the Parameters in a Regression Model." The degrees of freedom for SST, SSR, and SSE given above follow from his guideline that the degrees of freedom in a regression model are "the number of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation."

The standard error of the estimate for this bivariate regression model, s_e, is found by taking the square root of MSE, the mean square error.

$$s_e = \sqrt{MSE} = \sqrt{1.95} = 1.396$$

The standard error of the estimate is used to estimate the confidence band around the regression line as shown in the figures below. The default confidence band in the type of Stata graphs shown below is the 95% confidence interval band. This means that approximately 95% of the bivariate datapoints in the regression analysis are expected to be within the confidence band.

A4.3 Calculating Bivariate Regression Using Stata

The regression model for Dataset C is run in Stata 18.5 to compare results to those obtained in the numerical calculations just demonstrated. To do this, enter the paired X and Y values for Dataset C into a spreadsheet such as Excel, then copy and paste them into the "Data Editor" spreadsheet of Stata. Next, the command "**reg y x**" is entered into the command line of Stata to run the bivariate regression model for these data. The output of Table A4.1 is obtained.

In the upper left of the table, the fundamental ANOVA results are given. They verify the values obtained with the demonstration calculations for Dataset C obtained in this appendix: $MSR = 72.2$; $MSE = 1.95$; $MST = 16$. The summary statistics accompanying these ANOVA results are displayed in the upper right quadrant. There were only six observations from which an F ratio of 37.03 was obtained, with 1 and 4 degrees of freedom, significant beyond the 0.01 level, with an R^2 value of 0.9025, and a standard error of the estimate of 1.396. The standard error is labeled as "Root MSE," which stands for "square root of the Mean Squared Error value." This agrees with the standard error calculation shown above and was calculated in the same way, by taking the square root of the MSE.

The following command can be used to obtain a simple bivariate scatterplot as shown in Figure A4.3, which also corresponds well to the graph of Dataset C in Figure A4.1, which was created using Excel.

Table A4.1 Stata output table summarizing the results of the bivariate regression analysis of Dataset C.

. reg y x						
Source	**SS**	***df***	**MS**		**Number of obs. = 6**	
					$F(1, 4) = 37.03$	
Model	72.2	1	72.2		Prob > $F = 0.0037$	
Residual	7.8	4	1.95		$R^2 = 0.9025$	
					Adj. $R^2 = 0.8781$	
Total	80	5		16	Root MSE = 1.3964	
y	Coeff.	Std. err.	t	$P > \|t\|$	[95% conf. interval]	
x	1.9	.3122499	6.08	0.004	1.033055	2.766945
_cons	0.5	1.662077	0.30	0.779	−4.114666	5.114666

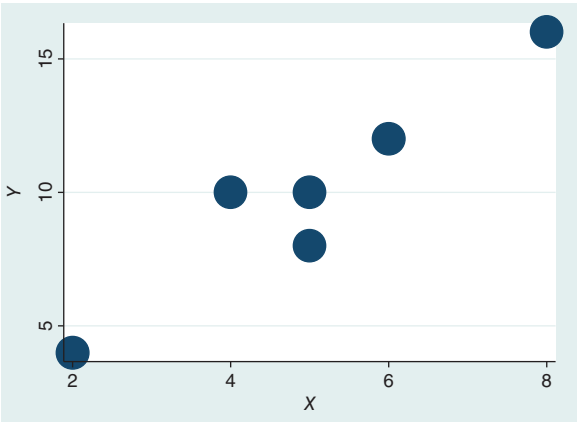


Figure A4.3 A simple bivariate scatterplot for the bivariate regression analysis of Dataset C using the "**graph twoway scatter**" command of Stata.

graph twoway scatter y x, msize(7)

There are three parts to this Stata command: (1) construct a ***graph***; (2) make it a ***twoway*** graph; and (3) use the ***scatter y x*** option because we wish to have a scatterplot of the relationship between *X* and *Y* with ***y*** (the criterion) as the vertical axis and ***x*** (the predictor) as the horizontal axis. The "msize(huge)" option is given after the comma to request large data points on the graph.

A slightly improved version of this command will return the same bivariate scatterplot, but with a regression line drawn in and the 95% confidence interval band around the regression line as shown in Figure A4.4.

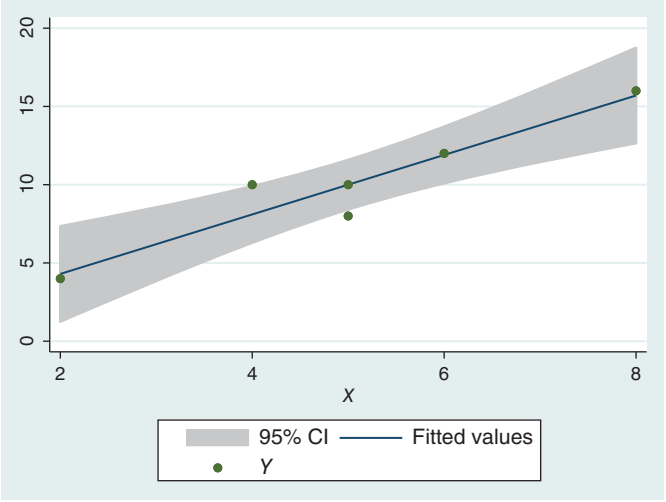


Figure A4.4 An improved graphic for the regression analysis of Dataset C by using additional options to add a regression line and a 95% confidence interval band to the bivariate scatterplot of Figure A4.4.

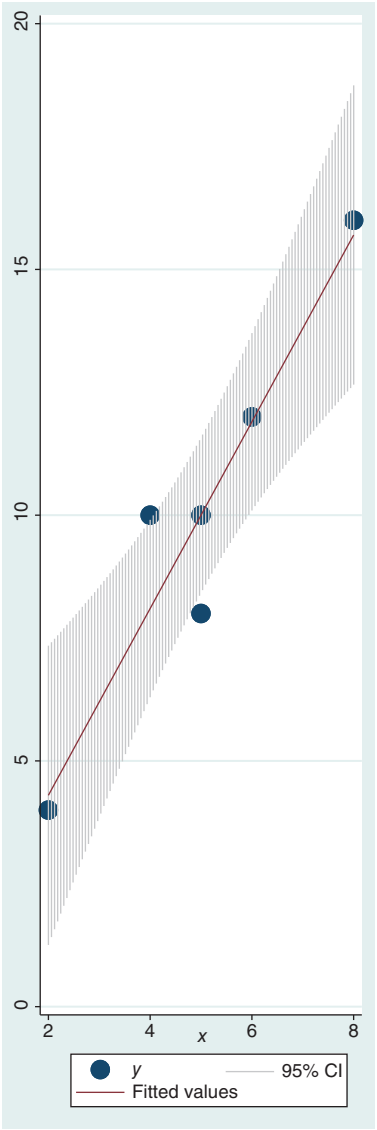


Figure A4.5 A third iteration of the bivariate scatterplot graphic that accompanies the screencapture of the Stata bivariate regression analysis output shown in Figure A4.3. This iteration adds to the regression line and 95% confidence interval (which was created in Figure A4.4), a more accurate aspect ratio to better represent the true regression line slope and the shape of the confidence interval band.

graph twoway (lfitci y x)(scatter y x)

The "lfitci" option within the first set of parentheses indicates ("lf" = line fit, and "ci" = with a confidence interval). This particular scatterplot graph is somewhat misleading because it does not have the same aspect ratio as the dataset that went into it. The vertical variable ***y*** had a standard deviation of 4 and the horizontal variable ***x*** had a standard deviation of 2. So, the correct aspect ratio should be two-to-one, which is twice as tall as wide, but this figure is closer to being a one-to-two aspect ratio.

A more complex command is used to obtain a bivariate scatterplot with a more accurate aspect ratio between the *Y* and the *X* axes, while also having the regression line and the confidence interval added to the scatterplot as shown in Figure A4.5.

graph twoway scatter y x, xsize(4) ysize(12) msize(huge) || lfitci y x, ciplot(rspike)

Notice that the commands "xsize(5)" and "ysize(10)" create the proper aspect ratio of two-to-one, with twice as much standard deviation for the ***y*** variable as for the ***x*** variable. Again, the "lfitci" option creates a regression line with a confidence interval. The confidence interval band is directed to be a lighter-shaded band by the "ciplot(rspike)" instruction. The entire command creates the scatterplot shown in Figure A4.5.

A4.4 Study Questions for Appendix 4

Bivariate Regression

Essay Questions

1. What is the "criterion" variable in bivariate regression? What is the "predictor" variable? Which one is said to be "regressed onto the other?
2. What is a regression line and how it is used?
3. What is another name for the "mean square" statistic?

4 What is another name for the "sum of squares" statistic?
5 What is the matrix formula for a CSSCP matrix? What does each entry in that matrix consist of (both the diagonal entries, and also the off-diagonal entries)? What is the matrix formula for an SSCP matrix and what does each entry consist of?
6 What is the matrix formula for a covariance matrix? What does each entry in that matrix consist of (both the diagonal entries, and also the off-diagonal entries)?
7 What statistic appears in the diagonal entries of a covariance matrix? What statistic appears in the off-diagonal entries of a covariance matrix?
8 How is the Pearson product moment correlation coefficient (PPMC) obtained from the covariance?
9 What do we mean when we say that the PPMCC is "standardized covariance"? How does one standardize a variable? How does one standardize a covariance statistic?
10 What is a coefficient of determination? How is it calculated? How is it interpreted?
11 What is the equation for obtaining the slope, **b**, in bivariate regression? Which standard deviation is the numerator of the ratio in the equation? Why? Which is the denominator? Why?
Once you have the slope of the regression line calculated, how can you obtain the *Y* intercept?
12 How does one write the regression equation? What is the "*Y*-hat" symbol in that equation, and what does it represent?
13 What do we mean when we say that the regression line is a "least squares" line? What other statistical quantity do we refer to as a "least squares" statistic? How are these two statistics similar? In what ways do they differ?
14 At what value of *X*, the predictor variable, does the regression line cross the *Y* axis? What is the *Y* value at that point referred to as?
15 What is the quantity *SSR*, and how is it obtained? How is it related to the coefficient of determination?
16 What is the quantity *SSE*, and how is it obtained? How is it related to the coefficient of nondetermination?
17 In one of the final paragraphs of Appendix 4, it states that "...in bivariate regression, it is actually the *variation* quantities (the sums of squares) rather than the *variance* quantities that best illustrate the additivity of predicted and error partitions." Explain.

Calculation Questions

For each of the following questions, you may do the calculations either with a pocket calculator or using the matrix capabilities of Excel. Use Stata for the creation of the scatterplots with their regression lines and their confidence intervals.

1 In Appendix 4, it was demonstrated how matrix algebra can be used to calculate a covariance matrix, and how a correlation coefficient can be calculated from the resulting information. This was demonstrated for Dataset C, with an $r = 0.95$. Now do that same analysis for Dataset B.
2 Now repeat that same correlation coefficient calculation analysis process again, but in this third calculation using Dataset A.
3 In Appendix 4, a "correlational fourfold" summary table was created for Dataset C. Construct similar fourfold tables for Datasets B and A.
4 Calculate the slope and the *Y* intercept for Datasets B and A in the same manner demonstrated in Appendix 4 for Dataset C, and create the regression equation for each of these datasets.

5 In Appendix 4 (Eq. (A4.5)), it was demonstrated that the coefficient of determination is indeed equal to the sum of squares regression, SSR, divided by the total sum of squares on *Y*, SST. Demonstrate that the same is true of Datasets B and A.

6 In Eq. (A4.8) of Appendix 4, it was demonstrated, using Dataset C, that the total sum of squares on the criterion variable *Y*, can be partitioned into two additive parts: the predictable part which is the sum of squares regression (SSR), and the error or unpredictable part which is the sum of squares error (SSE). Do that same demonstration with Datasets B and A.

7 In Eq. (A4.9) of Appendix 4, it was demonstrated using Dataset C that the coefficient of nondetermination is equal to the sum of squared errors, SSE, divided by the total sum of squares, SST. Demonstrate that the same is true of Datasets B and A.

8 In footnote 3 of Appendix 4, it was stated that "the regression line is a least squares line in the sense that the sum of squared deviations of observed *Y* scores from the corresponding predicted *Y* scores on the regression line are less than from any other line." To demonstrate that this is true, numerically "superimpose" (or combine) the more steep regression line from Dataset C with the more shallowly sloped bivariate data points from Scatterplot B. Then, show that the sum of squared differences for this more steep line, as deviations from the actual points, is greater than the sum of squared differences for Dataset B when you use its own fitted regression line.

9 Run the Stata command for a scatterplot with its regression line and confidence interval that was given at the end of Appendix 4, but this time with Datasets B and A, and compare them to the scatterplot that was given for Dataset C. What do you learn about predictive accuracy from the comparison of the scatterplots, regression lines, and confidence intervals for these three simple datasets, with $r = 0.45$ compared to $r = 0.75$ and to $r = 0.95$?

To the psychometrics student: Thank you for your diligence and hard work. You will be rewarded with increased skill and insight into the regression analysis process.

Appendix 5

Multiple Regression Analysis using Matrix Algebra and Stata

APPENDIX MENU

A5.1 The Graphical Basis of Multiple Regression, 538
A5.2 Computational Steps Underlying Multiple Regression, 545
- A5.2.1 Step 1: Calculate the Correlation Matrix, 545
- A5.2.2 Step 2, Part 1. Calculate the Multiple (Two or More) Slopes, then the *Y*-Intercept, and then Write the Predictive Equation, 547
  - A5.2.2.1 Part 1. Partition the Correlation Matrix, 547
  - A5.2.2.2 Part 2. Obtain the Matrix Inverse of the R_{xx} Partition, 548
  - A5.2.2.3 Part 3. Calculate the Standardized Regression Coefficients as the Product of R_{xx}^{-1} and r_{xy}, 548
  - A5.2.2.4 Part 4. Calculate the Unstandardized Regression Coefficients by Multiplying the Standardized Coefficients by the Ratios of Their Respective s_y and s_x Values, 549
  - A5.2.2.5 Part 5. Solve for the *Y*-intercept, 549
  - A5.2.2.6 Write the Multiple Regression Predictive Equation, 550
- A5.2.3 Step 3. Create Predicted Scores on *Y* from Each of the Seven Pairs of X_1 and X_2 Scores, 550
- A5.2.4 Step 4. Calculate SSR, the Sum of Squares Regression, 550
- A5.2.5 Step 5. Calculate the Errors of Prediction by Subtracting Predicted Scores, $\hat{Y}$, from Actual *Y* Scores, 551
- A5.2.6 Calculate a Simple Structure, that is, *a Clear and Interpretable Factor Solution*. However, SSE, the Sum of Squares Error, 551
- A5.2.7 Step 7. Confirm that $SST = SSR + SSE$, and also that the Squared Multiple Correlation Is Equal to $R^2 = SSR/SST$, and that One Minus the Squared Multiple Correlation Is Equal to $1 - R^2 = SSE/SST$, 552
- A5.2.8 Step 8. Calculate the ANOVA Table Statistics: The Three Mean Squares (MSR, MSE, and MST), the *F* Ratio, and the Standard Error of the Estimate, $s_{y.x}$, 552

A5.3 Multiple Regression Results with a Reasonably Larger Sample, 553
A5.4 Discussion, 555
A5.5 Study Questions for Appendix 5, 556

Multiple regression is the generalization and expansion of the bivariate linear regression model, which is demonstrated in Appendix 4. Whereas *bivariate* regression deals with only a single predictor variable *X* that is used to predict a single criterion variable *Y*, *multiple regression analysis* expands this process to predict a single criterion variable *Y* from two or more predictor variables.

Predicting the criterion variable *Y* from multiple combined variables does more than just compare the several predictors in their predictive power. It can often be an example of the Gestalt

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

psychology mantra that "the whole is greater than the sum of its parts," but in other ways, it is an example of the whole being less than the sum of its parts. With regard to the percentage of variance in the criterion variable accounted for by the predictors, the whole with two or more predictors is never more than and almost always less than the sum of the parts. However, multiple regression analysis between a criterion variable and multiple predictors provides substantially more insight into the nature of the predictive relationships than the results of separate bivariate regression analyses from those same predictors, as will be shown in what follows. As will also be shown, the percent of variance accounted for, the R^2 value, from a set of X variables combined in a multiple regression, can often be substantially larger than the R^2 value from any of them as a single predictor. The R^2 value from multiple combined predictors will always be *at least as great* as the largest R^2 from any one of them separately.

The principles and rationale of multiple regression will be illustrated geometrically and mathematically in this appendix in its simplest and most easily understood case—that of predicting a Y variable from only two X variables. Since this simplest case involves only three variables, one criterion variable Y, and the two predictor variables X_1 and X_2, we will be able to demonstrate multiple regression geometrically within a three-dimensional space, a space that can be easily visualized. This three-variable case of the multiple regression model is defined as "regressing the criterion variable Y onto two predictors, X_1 and X_2." The method referred to as regression analysis is diverse and quite expandable. It comes in many flavors. It can model a large number of predictors, with or without interactions among those predictors, and even multiple criterion variables (so-called "multivariate multiple regression").

A story problem will be used to illustrate the example calculations. Suppose in a science fiction movie there are many starship battle cruisers in a fleet, each having a crew of eight persons. The crew members all come from one of two planets, Cancri or Gleise. The Gleiseans have been the dominant force in the star fleet and have often been known to harass the Cancrians, who numerically still constitute a minority in most crews. Harassment is forbidden in the starship fleet but still occurs. The fleet commander commissions a study to identify "risk factors" of harassment, and then to identify the strongest predictor of harassment, the major cause. They have narrowed it down to two competing causes, numerical dominance of the Gleiseans in many starship crews, making Cancrians the picked-on minority; and secondarily, an aggressively competitive interpersonal climate in some of the crews. They randomly sample seven starships from the fleet, keep a record of harassment incidents over a 15-week period (Y), and administer an "aggressively competitive climate" questionnaire (X_1) to each crew member.

The dataset from the sample of seven starships is shown below with Y, the number of harassment incidents during the 15 weeks measured, ranging from 4 for Starship F to 12 for Starship D, as shown in Table A5.1. Climate, variable X_1, is measured on a 12-point competitive-climate scale administered to each crew member, with a score of 12 indicating that the climate within the starship crew is maximally competitive. The numerical dominance of Gleiseans, variable X_2, ranges from a score of only one out of the eight crew members (12.5% dominance) being Gleiseans in Starship F, to Starship A, where seven out of the eight crew members are Gleiseans (87.5% Gleisean dominance).

The first steps in demonstrating the geometrical, conceptual, and mathematical basis of multiple regression consist of calculating the variances and the covariances of these three variables Y, X_1, and X_2, and then obtaining the correlation matrix showing how each variable correlates with the other two. These calculations will be accomplished using Stata. First copy and paste into Stata's data editor, row 1 of Table A5.1, which contains the labels of the variables, Y, X_1, and X_2. Next copy and paste the seven rows of data (not including the Starship labels on the left). When Stata inquires

Table A5.1 Labeled data table of the seven starships with criterion variable Y and predictor variables X_1 and X_2.

	Harassment Incidents	Climate	Numerical dominance
	Y	X_1	X_2
Starship A	9	5	7
Starship B	10	1	6
Starship C	7	2	4
Starship D	12	11	4
Starship E	8	9	3
Starship F	4	4	3
Starship G	6	10	1

whether you would like the first row to be variable names, select the affirmative. Table A5.1 facilitates transfer of the data in electronic form from a pdf or ebook format.

	y	x1	x2	
1	9	5	7	
2	10	1	6	
3	7	2	4	
4	12	11	4	
5	8	9	3	
6	4	4	3	
7	6	10	1	

Next, enter the "**correl y x1 x2, covariance**" command into Stata (that is, the correl command with the covariance option after the comma) in order to create the variance/covariance matrix (see Appendix 2), with the three variances in the diagonal entries and the three covariances in the triangle of values below the diagonal. In the output table, Y has a variance of 7, X_1 has a variance of 16, and X_2 has a variance of 4. We could make a preliminary observation that the competitive climate predictor has a substantially higher variance than does the numerical dominance of Gleiseans.

```
. correl y x1 x2, covariance
(obs=7)

             |        y        x1        x2
-------------+---------------------------
           y |        7
          x1 |  2.16667        16
          x2 |  2.83333 -4.33333         4
```

Next obtain the correlation matrix using the **correl** command (this time with the "covariance" option omitted). Clearly, X_2 is the stronger of the two predictors of harassment (Y), with a correlation coefficient of 0.5354 compared to 0.2047 for the X_1 predictor. Predictors X_1 and X_2 are also negatively correlated with one another, $r = -0.5417$.

```
. correl y x1 x2
(obs=7)

             |        y       x1       x2
-------------+---------------------------
           y |   1.0000
          x1 |   0.2047   1.0000
          x2 |   0.5354  -0.5417   1.0000
```

A5.1 The Graphical Basis of Multiple Regression

The negative correlation between X_1 and X_2 can be visualized using the Stata command "**graph twoway scatter**" to create a scatterplot showing the correlative structure of these two predictor variables. The left half of Figure A5.1 is the scatterplot as created by Stata. The same scatterplot is repeated on the right but this time enhanced by grid lines, coordinates, and starship labels, added to the Stata plot for clarification.

```
. graph twoway scatter x2 x1, msize(4) xsize(13) ysize(10)
```

The inverse relationship indicated by the negative correlation coefficient between the two predictors can be visualized in Figure A5.1. The linear trend of the data points in that figure is from upper left down to lower right, rather than increasing from lower left up to upper right as would be expected for a positive correlation. This scatterplot pattern reveals a tendency for those starships with relatively high competitiveness (E, G, and D) to be markedly lower in Gleisean numerical dominance. A moderate negative correlation like this between the predictors indicates that the predictive value contributed by one predictor is not redundant with the predictive value contributed by

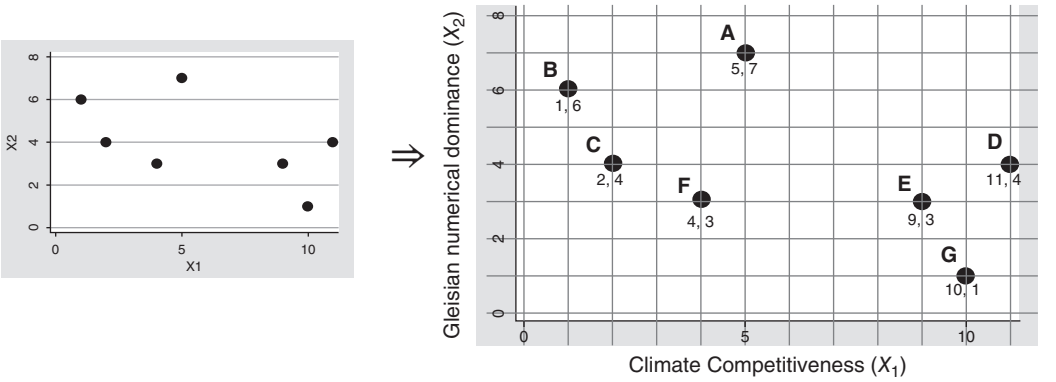
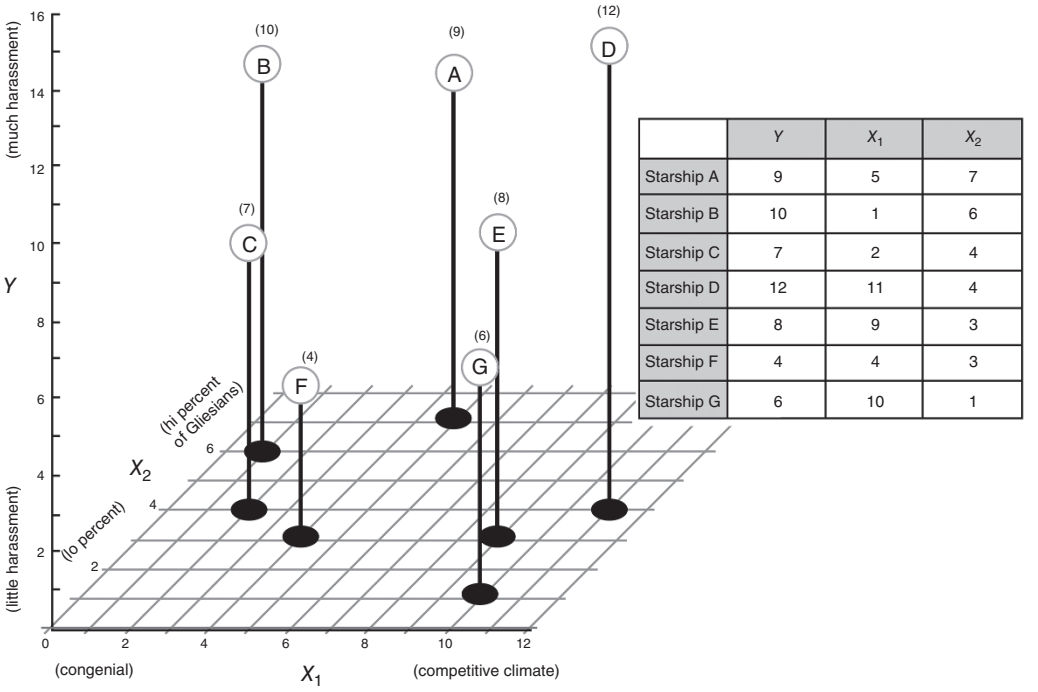


Figure A5.1 Standard Stata scatterplot of X_1 and X_2 (*left*), enhanced with labels and coordinates (*right*).



	Y	X_1	X_2
Starship A	9	5	7
Starship B	10	1	6
Starship C	7	2	4
Starship D	12	11	4
Starship E	8	9	3
Starship F	4	4	3
Starship G	6	10	1

Figure A5.2 Three-dimensional scatterplot showing how the position of each starship in the X_1 by X_2 grid accounts for its score on the Y dimension, harassment.

the other. Under these circumstances, one would expect the multiple regression analysis to have a relatively higher R^2 value than otherwise.

Figure A5.2, a three-dimensional scatterplot, is built upon the two-dimensional 11×7 scatterplot grid formed by the axes of the two predictors, X_1 and X_2. A vertical riser extends from each of the seven X_1 by X_2 starship locations on this grid to a small circular marker at the top labeled A through G, one for each of the starships. This vertical dimension, the Y axis, is the criterion variable, the number of harassment incidents. The height of each riser in this vertical dimension indicates the number of harassment incidents for each of the seven starships. This indicates how harassment is dependent upon or predicted from each starship's location on the grid of the two predictors, competitive climate (X_1) and numerical dominance (X_2). An inset of the original dataset table has been placed in Figure A5.2 in close proximity to the three-dimensional scatterplot to facilitate comparisons of the scores of each starship on the predictors and on the harassment score Y. These relationships hold the answer to the question of how well the two risk factors, X_1 and X_2 can predict harassment.

Now that the graphical representation of the geometrical relationship among these three variables has been constructed, a picture of it can be taken from the front, and a picture from the right. The two pictures are compared, as shown in Figure A5.3. The frontal view picture, harassment (Y) as a function of competitive climate (X_1) is created in Stata by the first command shown below. The picture from the right side, harassment (Y) as a function of Gleisean numerical dominance (X_2), is created using the second Stata command given below.

```
. graph twoway scatter y x1, msize(5) xsize(10) ysize(10) || lfit y x1
. graph twoway scatter y x2, msize(6) xsize(8.5) ysize(10) || lfit y x2
```

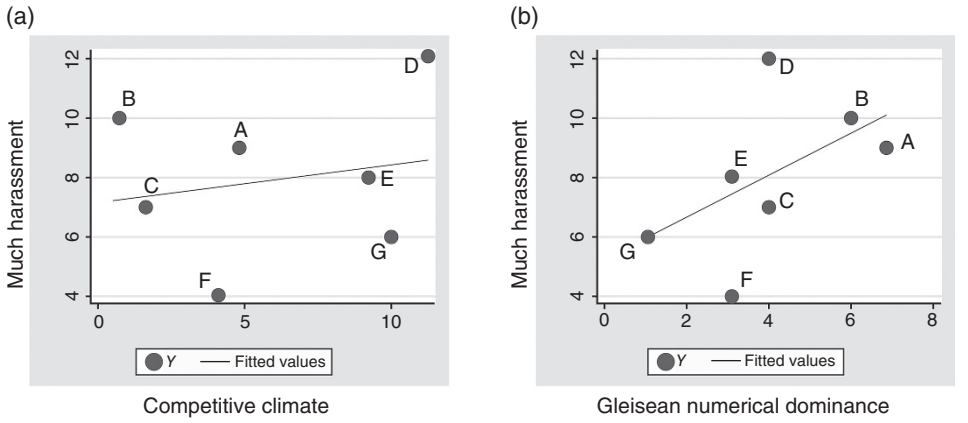


Figure A5.3 Two bivariate scatterplots with the one on the left showing the prediction of harassment from climate, and the one on the right showing the prediction of harassment from numerical dominance. The respective regression lines are superimposed on the scatterplots. (a) Frontal view $Y = f(X_1)$ and (b) right side view $Y = f(X_2)$.

As can be clearly seen from these two scatterplots (and the correlation matrix in the introduction to this appendix), Gleisean numerical dominance (X_2) is a substantially stronger predictor of harassment than competitiveness of the starship climate (X_1). The line of X_2 has a markedly steeper predictive slope and less scatter of the datapoints around the line. The line through each of the two sets of bivariate points is the regression line, with the slope calculated in the manner demonstrated in Appendix 4, Bivariate Regression. The Stata command for computing the bivariate regression between numerical dominance (X_2) and harassment (Y) is "**regress y x2**." From the Stata output for the X_2 predictor shown below, it can be seen that the regression line for this analysis is $\hat{Y} = 0.708\,X_2 + 5.167$, with a slope of approximately 0.708 and a Y-intercept of approximately 5.167. That is, in the coefficients column ("Coef.") of the regression output table, the Y-intercept ("_cons") for the regression line shown in the scatterplot on the right is 5.166667 (5.167 to three places), and the slope (rise-over run) is 0.7083333 (0.708 to three places). That means that for every one-unit increase in the number of Gleiseans among the crew of eight, there is a 0.708 unit increase in harassment, which creates an angle of incline of about 35°.

Besides the slope and the intercept, the F ratio and its probability (given in the upper right corner of the output), and also the R^2 value given right below them, also serve as indicators of the strength of the predictive relationship. The F ratio of 2.01 is not very large, nor is its probability value, $p = 0.2155$, statistically significant (which would require that it be less than 0.05), but given the unusually small sample size, that is not surprising. If the number of observations were doubled from 7 to 14, the F ratio would be fairly strong and statistically significant, as shall be seen when we do so in part 3 of this appendix (A5.3). The R^2 value indicates that the proportion of variance accounted for is 0.2867 (about 28.7%), and the adjusted R^2 ("Adj R-squared")[1] is only about half of that, 0.1440 (about 14.4%).

1 The adjusted R^2 is the R^2 value adjusted for its degrees of freedom, that is, the R^2 value adjusted for the number of parameters involved in making the prediction.

Bivariate Regression Analysis for Y = bX_2 + a

```
. regress y x2
```

Source	SS	df	MS		
Model	12.0416667	1	12.0416667	Number of obs	= 7
Residual	29.9583333	5	5.99166667	F(1,5)	= 2.01
				Prob > F	= 0.2155
Total	42	6	7	R-squared	= 0.2867
				Adj R-squared	= 0.1440
				Root MSE	= 2.4478

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x2	.7083333	.4996527	1.42	0.215	-.5760647	1.992731
_cons	5.166667	2.202362	2.35	0.066	-.4946842	10.82802

The Stata output from the X_2 prediction of harassment was presented first because it is the stronger predictive relationship. Now the Stata output for the prediction of harassment from X_1, the weaker predictor, shall be considered. The Stata command is "**regress y x1**" or, in short form as shown here, it is "**reg y x1**." Either will work. The regression line for the analysis of the prediction of harassment from competitiveness of the climate, X_1, is $\hat{Y} = 0.135\,X_1 + 7.188$. That is, the *Y*-intercept for the regression line shown in the scatterplot of the left panel of Figure A5.3 is approximately 7.188, and the slope is approximately 0.135, a rise of only 0.135 harassment units for every one-unit change in climate. The angle of incline is about 7½ degrees (arctan [0.135] = arctan[rise/run] = 7.69°). The R^2 value (0.0419) is also quite small for this weaker predictor.

Bivariate Regression Analysis for Y = bX_1 + a

```
. reg y x1
```

Source	SS	df	MS		
Model	1.76041667	1	1.76041667	Number of obs	= 7
Residual	40.2395833	5	8.04791667	F(1, 5)	= 0.22
				Prob > F	= 0.6597
Total	42	6	7	R-squared	= 0.0419
				Adj R-squared	= -0.1497
				Root MSE	= 2.8369

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	.1354167	.2895384	0.47	0.660	-.6088654	.8796987
_cons	7.1875	2.041487	3.52	0.017	1.939689	12.43531

The two Stata outputs just presented show how much can be done with bivariate regression, the simpler regression analysis method demonstrated in Appendix 4, where the *Y* variable can be predicted separately from each one of the two *X* variables using a two-dimensional scatterplot and a single regression line. Much additional insight can now be gained by moving up from these two separate bivariate regression analyses to a single holistic multiple regression analysis whereby harassment, *Y*, can be predicted from the combination of the two X variables simultaneously. Rather than a two-dimensional scatterplot, like the two shown in Figure A5.3, the results can be visualized using the three-dimensional scatterplot shown in Figure A5.2, and rather than using a two-dimensional regression line as shown in each of the two bivariate regression analyses of Figure A5.3, a three-dimensional regression plane can be constructed.

Multiple Regression Analysis for $\hat{Y}$=.463X_1 + 1.210X_2 +.381

```
. regress y x1 x2

      Source |       SS       df       MS              Number of obs =       7
-------------+------------------------------           F(2, 4)       =    3.45
       Model |  26.5921376     2  13.2960688            Prob > F      =  0.1346
    Residual |  15.4078624     4   3.8519656            R-squared     =  0.6331
-------------+------------------------------           Adj R-squared =  0.4497
       Total |        42       6           7            Root MSE      =  1.9626

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    .463145   .2382976     1.94   0.124    -.1984752    1.124765
          x2 |    1.210074   .4765951     2.54   0.064    -.1131666    2.533314
       _cons |    .3808354   3.030134     0.13   0.906    -8.032165    8.793836
------------------------------------------------------------------------------
```

The three-dimensional scatterplot in Figure A5.2 illustrates well the central concept of multiple regression. *It is the prediction of a single criterion variable* (*Y*), *by two or more multiple predictors* (X_1 and X_2 in this case), *using a regression plane rather than regression lines*. If you were to take a perfectly flat plane and try to give it the best fit to all seven of the starship locations in the 3D scatterplot of Figure A5.2, clearly that plane would have a strong slope upward as it goes to the back of the graphic space where starship B, A, and D are found (having high *Y* harassment values of 10, 9, and 12, respectively), and a slope downward in the front of the space with starships F and G (having lower *y* values of 4 and 6). This is a least-squares plane, a regression plane, just like the line in bivariate regression is a least-squares line. This plane has less squared deviation from the actual data points than any other plane.

The Stata command for multiple regression is simple enough, "**regress y x1 x2**." It is very similar to the bivariate regression command. Stata's multiple regression output table also closely resembles the one for bivariate regression. The focus will first be on the intercept and the two slopes in this table of results. These three values are circled on the Stata output, with $\mathbf{b_1} = 0.463145$ being the slope for the X_1 predictor, $\mathbf{b_2} = 1.210074$ being the slope for the X_2 predictor, and $\mathbf{a} = 0.3808354$ being the *Y*-intercept. Looking now at Figure A5.4, the intercept and the two slopes can be used to reveal two sides of the regression plane. In front, the first side of the predictive plane rises up and to the right above the X_1 axis. It does indeed have a slope ("rise-over run") of 0.463145. For each unit to the right along the X_1 axis, the line goes up approximately 0.463 units vertically in a *Y* direction. In other words, variable X_1, competitiveness, appears to have at least some predictive power in predicting *Y*, harassment. But X_2, numerical dominance, has more predictive power with a slope of 1.210. It will therefore have more effect on the location of the plane that will be fitted to these two sides, which is added to the graph of Figure A5.4 to create Figure A5.5. Notice that, with a *Y*-intercept value of $\mathbf{a} = 0.381$, the starting point of the plane on the lower left where it intersects with the *Y*-axis is just a small amount higher than zero. The predictive plane extends upward from there.

This flat and straight predictive plane is shown in Figure A5.5. It maintains its linear integrity as it tilts upward on the right with a rise-over-run slope of 0.463. It is then tilted upward nearly three times that much from the front of the figure to the back to fit the plane to the slope rising up from axis X_2, which has a rise-over run of 1.210. This is, in fact, the geometry indicated by the multiple regression results received from the Stata output given above.

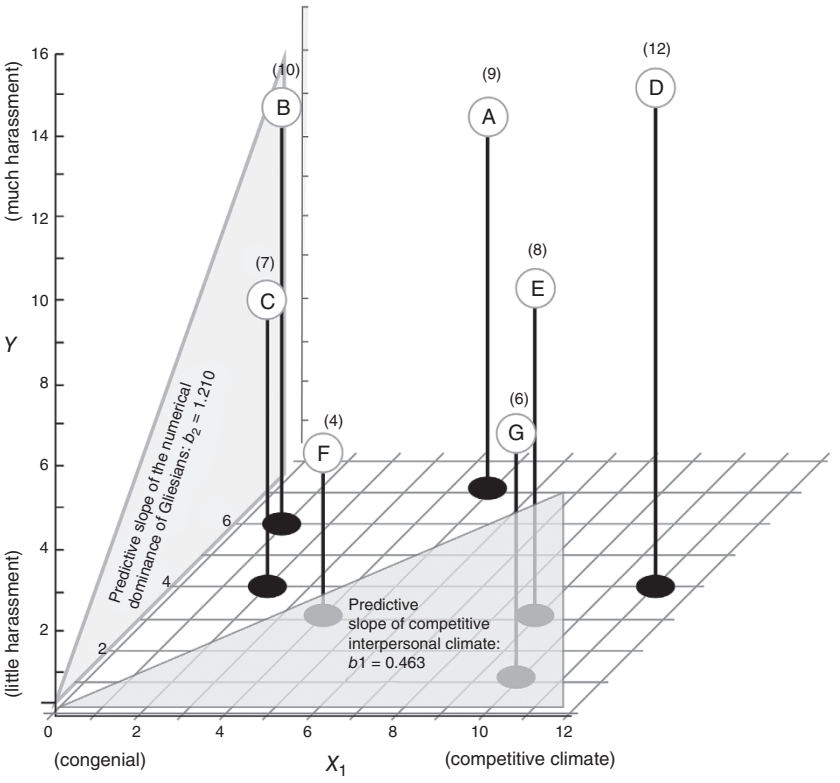


Figure A5.4 *Y*-intercept and the two slopes for predicting from X_1 and X_2 using multiple regression.

Now that the multiple regression plane has been constructed in Figure A5.5, with a little help from Stata, it can be observed that this is indeed a least-squares plane. The sum of the squared deviations (i.e. errors) above and below the plane is less than for any other plane. Notice that the deviations above the plane, where the actual *Y* value is higher than the value predicted by the plane, are shown with bold numerical values, while those below the plane are grayed by the semi-transparent plane. Reading off the values in ascending order from starship A to starship G, they are: −2.167, **1.896**, **0.853**, **1.684**, −0.179, −1.864, and −0.222. This multiple regression predictive plane is, of course, analogous to the bivariate regression predictive line in that the sum of the values above and below it equals zero. The sum of the positive values above the plane is **1.896 + 0.853 + 1.684 = 4.432**. The sum of the negative values below the plane is (−2.167) + (−0.179) + (−1.864) + (−0.222) = −4.432. The combined sum of deviations above and below the plane is therefore zero.

The multiple regression plane can now be used to demonstrate two things. First, since this is linear regression, it has a linear relationship between the combined two predictors, X_1 and X_2, and the criterion variable *Y*. The multiple regression plane is linear and, therefore, perfectly flat, just as a bivariate regression line is linear and perfectly straight, having no curvature. The second thing demonstrated is that the linear relationship is additive. The plane is constructed by starting at the *Y*-intercept. This is the location where the plane intersects the *Y*-axis, analogous to the *Y*-intercept in bivariate regression. Added to that starting point are the numerical effects of the upward or downward slopes of each of the predictors.

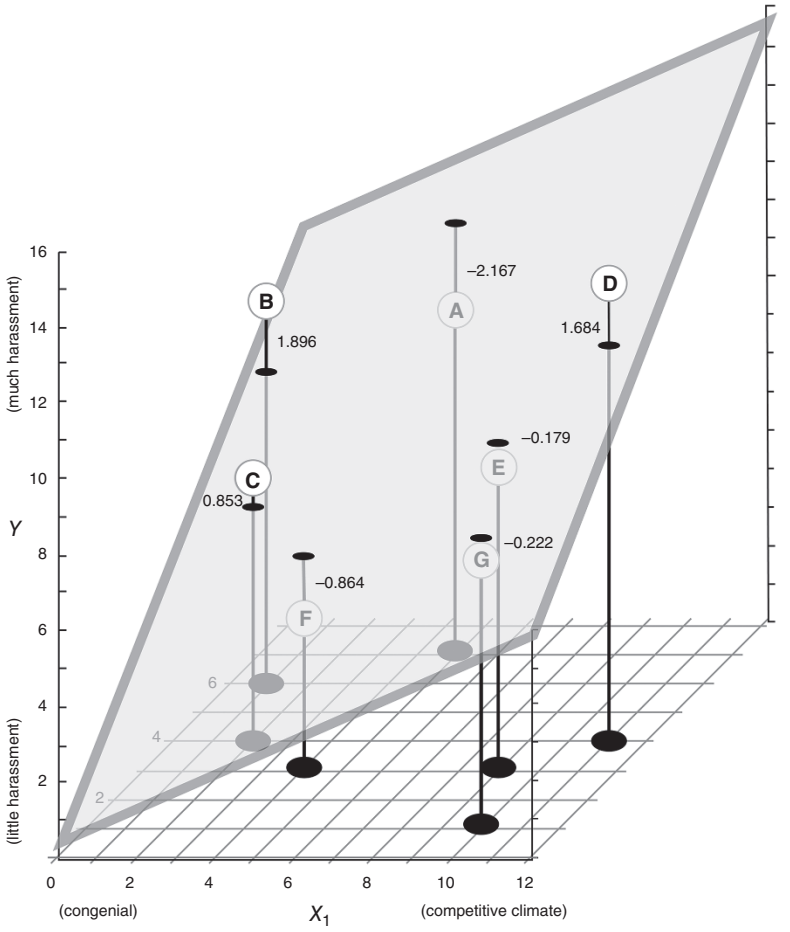


Figure A5.5 Three-dimensional regression plane for predicting Y from X_1 and X_2.

Given the two slopes, $b_1 = 0.463$ and $b_2 = 1.210$, and the Y-intercept, $a = 0.381$, labeled in the Stata results table as "cons," which is short for "constant," the regression equation for this dataset can now be precisely stated as $\hat{Y} = 0.463X_1 + 1.210X_2 + 0.381$. These predicted Y scores are obtained in Stata by using the "predict" command, **predict ypred**, or more generally **predict** [*newvar*], where the parenthetical "*newvar*" can be replaced by "**ypred**" or any other name the user chooses for the output variable.

A second Stata command, **generate errors = y – ypred**, will create the errors of prediction. The first word, "**generate**," is the Stata command that generates a new variable. It is here given the simple name of "errors." The rest of the command, "**= y – ypred**," indicates to Stata that this new variable is to be generated by subtracting the "**ypred**" variable from the "**y**" variable. Stata will record the values of these two output variables, to which the names "ypred" and "errors" were assigned, in the data editor spreadsheet to the right side of the three input variables that were originally entered as "y," "x1," and "x2." This table of the values for these five variables, three input variables, and two output variables, for each of the seven starship crews may be copied from the

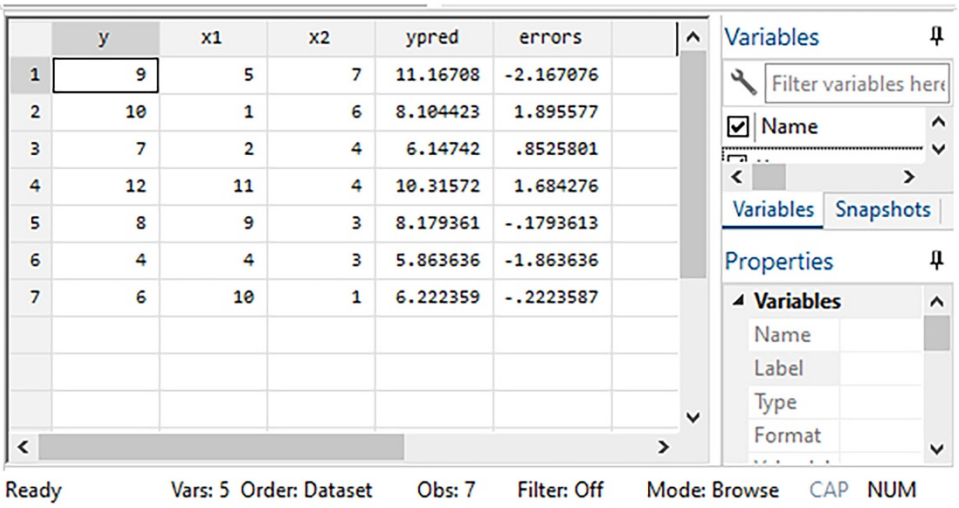


Figure A5.6 Screen capture of Stata's data spreadsheet with predicted scores and error scores appended to the right of the 7 × 3 matrix of the input data.

Stata data editor, shown in Figure A5.6 using the keystroke for copying (control-C for PC, or command-C for Apple).

Three of these columns are of particular interest for what was demonstrated in Figures A5.2 and A5.5. First, the Y score column (labeled "y") contains the criterion variable, the harassment scores for each of the seven starships. These seven scores give the heights of each of the vertical risers shown in Figure A5.2, on which the seven marker-spheres sit, with starship identity (A through G) written on each. Second, the column of $\hat{Y}$ predicted scores ("ypred") gives the height location where each starship's vertical riser intersects with the predictive plane, as shown in Figure A5.5. In other words, those intersections (marked by black ellipses) are the harassment scores as predicted by the regression plane. Third, the difference between actual harassment scores and predicted harassment scores, $Y - \hat{Y}$, in other words, the error scores (the "errors" column in Figure A5.5), give the distance that the harassment marker-spheres for each of the seven starships deviate from the predicted scores on the predictive plane. The positive error values above the plane are printed in bold digits, and the negative error values below the plane are grayed.

A5.2 Computational Steps Underlying Multiple Regression

Now that multiple regression geometry has been demonstrated, it is time to catch up on the mathematical calculations underlying that geometry. The following equations, formulas, and calculations are how these results would be obtained "by hand" if we did not have Stata or other statistical analysis applications to accomplish the calculations.

A5.2.1 Step 1: Calculate the Correlation Matrix

First, the correlation matrix is obtained for the three variables Y, X_1, and X_2. This begins with the calculation of the "centered" sum of squares and cross products matrix using the "hybrid" matrix

formula CSSCP = **X' x**, that is, the transpose of the raw data matrix **X**, post-multiplied by the deviation score matrix **x**.

$$\text{CSSCP} = \text{X}'\text{x} = \begin{bmatrix} 9 & 10 & 7 & 12 & 8 & 4 & 6 \\ 5 & 1 & 2 & 11 & 9 & 4 & 10 \\ 7 & 6 & 4 & 4 & 3 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 3 \\ 2 & -5 & 2 \\ -1 & -4 & 0 \\ 4 & 5 & 0 \\ 0 & 3 & -1 \\ -4 & -2 & -1 \\ -2 & 4 & -3 \end{bmatrix} = \begin{bmatrix} 42 & 13 & 17 \\ 13 & 96 & -26 \\ 17 & -26 & 24 \end{bmatrix}$$

Next, pre-multiply this matrix product CSSCP by the reciprocal of the degrees of freedom $\left(\frac{1}{N-1}\right)$ to obtain the variance/covariance matrix, **S**. This covariance matrix, calculated using the matrix functions of a spreadsheet, agrees with the values found above using the Stata command **correl y x1 x2, covariance**.

$$S = \begin{bmatrix} s_1^2 & s_{12} & s_{13} \\ s_{21} & s_2^2 & s_{23} \\ s_{31} & s_{32} & s_3^2 \end{bmatrix} = \left(\frac{1}{N-1}\right) X'x = \left(\frac{1}{6}\right) \begin{bmatrix} 42 & 13 & 17 \\ 13 & 96 & -26 \\ 17 & -26 & 24 \end{bmatrix} = \begin{bmatrix} 7 & 2.167 & 2.833 \\ 2.167 & 16 & -4.333 \\ 2.833 & -4.333 & 4 \end{bmatrix}$$

Next, create a standardizing matrix, *E*. It is a diagonalized matrix (meaning the off-diagonal elements are set to zero) with the diagonal elements set to the reciprocals of the respective standard deviations.

$$\boldsymbol{E} = \begin{bmatrix} 1/\sqrt{7} & 0 & 0 \\ 0 & 1/\sqrt{16} & 0 \\ 0 & 0 & 1/\sqrt{4} \end{bmatrix}$$

The covariance matrix *S* is now premultiplied by *E*, which standardizes it by rows (that is, divides each row by its standard deviation), and then the resulting *ES* product matrix is post-multiplied by *E*, which standardizes it by columns. The matrix formula expressing this process is $R = ESE$. In other words, the correlation matrix is obtained by pre-multiplying and post-multiplying the variance/covariance matrix *S* by *E*. This can be accomplished in Excel by the following command. **mmult(mmult(***the E matrix, the S matrix***),** *the E matrix*).

$$R = ESE = \begin{bmatrix} 0.378 & 0 & 0 \\ 0 & 0.250 & 0 \\ 0 & 0 & 0.500 \end{bmatrix} \times \begin{bmatrix} 7 & 2.167 & 2.883 \\ 2.167 & 16 & -4.333 \\ 2.883 & -4.333 & 4 \end{bmatrix}$$

$$\times \begin{bmatrix} 0.378 & 0 & 0 \\ 0 & 0.250 & 0 \\ 0 & 0 & 0.500 \end{bmatrix} = \begin{bmatrix} 1 & 0.205 & 0.535 \\ 0.205 & 1 & -0.542 \\ 0.535 & -0.542 & 1 \end{bmatrix}$$

A5.2.2 Step 2, Part 1. Calculate the Multiple (Two or More) Slopes, then the *Y*-Intercept, and then Write the Predictive Equation

Other than Step 1, where an entire correlation matrix was calculated rather than just a single correlation coefficient, this is the only step of the eight steps where multiple regression differs from bivariate regression. This second step for multiple regression is quite a big one. It moves the analysis to a higher level of mathematical sophistication with the inverse of a correlation matrix as the "computational engine." The inverse matrix takes into account the interrelationships of all of the predictors with one another. There are, of course, only two predictors in this simple example. In contrast, the computational engine for bivariate regression is just a single correlation coefficient. The unstandardized regression coefficient, that is, the slope of the regression line in bivariate regression, is created by multiplying the correlation coefficient by the ratio of the two standard deviations: $b = r\left(\frac{s_y}{s_x}\right)$. The second part of Step 2 for bivariate regression is also simple, calculating the *Y*-intercept by the equation: $a = \overline{Y} - b\overline{X}$. These two parameters are combined to create the bivariate regression equation for the third part of Step 2. If the slope b were equal to 1.90 and the intercept a were 0.5, then the bivariate regression equation would be $\hat{Y}_i = 1.90X_i + 0.5$.

In multiple regression, Step 2 will be counted as a single step, to keep the process comparable to that for bivariate regression, but with the recognition that Step 2 is a more complicated step for multiple regression. It consists of six parts:

1) partitioning the correlation matrix found in Step 1,
2) obtaining the inverse of the R_{xx} partition of the correlation matrix,
3) calculating the *standardized regression coefficients* as the product of R_{xx}^{-1} and r_{xy},
4) transforming the *unstandardized regression coefficients* into *standardized regression coefficients*, which are the slopes of the predictive plane,
5) solving for the *Y*-intercept, and
6) writing the predictive equation.

This more complex Step 2 for multiple regression will now be demonstrated with this six-part computational interlude.

A5.2.2.1 Part 1. Partition the Correlation Matrix

The correlation matrix obtained in Step 1 is partitioned into four parts.

R = ESE =

1	0.2047	0.5354
0.2047	1	–0.5417
0.5354	–0.5417	1

=

$\mathbf{R}_{yy}$	$\mathbf{r}_{yx}$
$\mathbf{r}_{xy}$	$\mathbf{R}_{xx}$

The upper left partition, $\mathbf{R}_{yy}$, is the single scalar value of 1. That is, it is the correlation coefficient of variable *Y*, the criterion variable, with itself.[2] The upper right partition $\mathbf{r}_{yx}$ is the 1×2-row vector of the correlations of the criterion variable *Y* with its two predictors, X_1 and X_2. The lower left partition, $\mathbf{r}_{xy}$, is just the transpose of $\mathbf{r}_{yx}$, the reflected image of the two predictors' correlations with their criterion, but in column vector form as a 2×1 vector. The partition of special interest in this

2 This is essentially a place holder. There are *multivariate* multiple regression models, that is, applications of multiple regression where there are multiple *Y* variables as well as multiple *X* variables, in which case the R_{yy} partition becomes a matrix and the r_{yx} and r_{xy} partitions are transformed from being vectors, to being matrices with dimensionality adjusted to fit with the dimensions of the R_{YY} and the R_{xx} partitions.

process is the lower right one, the $\mathbf{R}_{xx}$ partition, which is the matrix to be inverted, the intercorrelation matrix among the predictors. In this case, there are only two predictors. So, it is a 2×2 matrix.

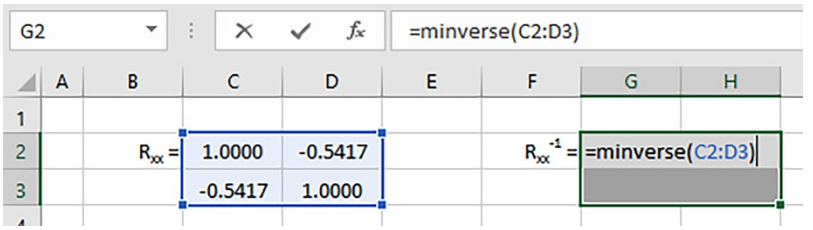
A5.2.2.2 Part 2. Obtain the Matrix Inverse of the R_{xx} Partition

In order to obtain the predictive coefficients for this multiple regression analysis, the inverse of the 2×2 R_{xx} matrix must be obtained. This can be accomplished in a spreadsheet such as Excel. To obtain the inverse, we place this R_{xx} matrix in the 2×2 matrix of cells (cells C2 and D2 in the second row of the spreadsheet, and cells C3 and D3 in the third row). We then enter the label "$R_{xx}^{-1} =$" in cell F2 in preparation for obtaining the inverse of the R_{xx} matrix.

	A	B	C	D	E	F	G	H	I
1									
2		R_{xx} =	1.0000	-0.5417		R_{xx}^{-1} =			
3			-0.5417	1.0000					

Now select the 2×2 group of cells from G2 to H3 (by dragging the cursor over them) to reserve the location for the resultant inverse matrix.

Enter the Excel command "=minverse(C2:D3)" while keeping that 2×2 location selected, then simultaneously holding down the "control" key and the "shift" key and pressing "enter."[3]



The inverse matrix $\mathbf{R}_{xx}^{-1}$ will appear in that selected location, as shown below.

	R_{xx} =	1.0000	-0.5417		R_{xx}^{-1} =	1.4153	0.7667
		-0.5417	1.0000			0.7667	1.4153

A5.2.2.3 Part 3. Calculate the Standardized Regression Coefficients as the Product of R_{xx}^{-1} and r_{xy}

To obtain the vector $\boldsymbol{\beta}$ of *standardized regression coefficients*, the R_{xx}^{-1} matrix is post-multiplied by the r_{xy} vector that was obtained by the partitioning operation above.

$$\beta = R_{xx}^{-1} r_{xy} = \begin{bmatrix} 1.4153 & 0.7667 \\ 0.7667 & 1.4153 \end{bmatrix} \begin{bmatrix} 0.2047 \\ 0.5357 \end{bmatrix} = \begin{bmatrix} 0.7002 \\ 0.9147 \end{bmatrix}$$

These two standardized regression coefficients in the $\boldsymbol{\beta}$ vector, $\beta_1 = 0.7002$ and $\beta_2 = 0.9147$ are like Pearson product-moment correlation coefficients in that both are in a standardized form, that is, they are shorn of their standard deviations. To utilize them for creating the regression equation, they must be converted into *unstandardized coefficients*, slopes if you will, analogous to the slopes defined in the bivariate regression equations of Appendix 4.

3 This three-key combination is an important one to learn and remember. It is the execution command for all matrix and array operations in the Microsoft Excel spreadsheet.

A5.2.2.4 Part 4. Calculate the Unstandardized Regression Coefficients by Multiplying the Standardized Coefficients by the Ratios of Their Respective s_y and s_x Values

In bivariate regression, there is only one Y variable and one X variable. So, the unstandardized regression coefficient, the slope, is $b = r\left(\frac{s_y}{s_x}\right)$. In multiple regression, there is a **B** vector of unstandardized regression coefficients where each of the slopes is the unstandardized regression coefficient multiplied by the ratio of the standard deviation of Y to the standard deviation of the appropriate X. The three standard deviations for this multiple regression dataset are $s_y = \sqrt{7} = 2.646$, $s_{x1} = \sqrt{16} = 4$, $s_{x2} = \sqrt{4} = 2$. The two slopes are therefore:

$$b_1 = \beta_1\left(\frac{s_y}{s_{x1}}\right) = 0.7002\left(\frac{2.646}{4}\right) = 0.7002\,(0.661) = 0.463$$

$$b_2 = \beta_2\left(\frac{s_y}{s_{x2}}\right) = 0.9147\left(\frac{2.646}{2}\right) = 0.9147\,(1.323) = 1.210$$

Notice that just as the correlation coefficient **r** in bivariate regression is converted into the slope **b** by multiplying it by the ratio of standard deviations, s_y/s_x, the same process is followed in multiple regression to convert the standardized regression coefficients into unstandardized regression coefficients, that is, slopes.

"Standardized" regression coefficients, like a Pearson correlation coefficient, express the magnitude of the relationship between two variables without reference to the standard deviation of either variable. On the other hand, in referring to the slopes, b_1 and b_2, as unstandardized regression coefficients, it means that each expresses the relationship between X and Y with reference to the ratio of their respective standard deviations. That is, these slopes convert raw X values from their own standard deviation units into a neutral form (by dividing by their own s_x), and then multiply that neutral form by s_y, which has the effect of expressing the predicted scores in the standard deviation units of Y. This process creates least-squares predicted scores ($\hat{\mathbf{Y}}$), that is, predicted scores in which the squared deviations between the predicted scores and the actual Y scores is minimized.

A5.2.2.5 Part 5. Solve for the *Y*-intercept

Also analogous to bivariate regression, the Y-intercept a can be obtained by entering the slopes and also the means of Y, X_1, and X_2 into the regression equation and solving for $\boldsymbol{a}$. The multiple regression equation is

$$Y' = b_1X_1 + b_2X_2 + a$$

To obtain the intercept a, the means of Y, X_1, and X_2 are entered into the general equation.

$$\overline{Y} = b_1\overline{X}_1 + b_2\overline{X}_2 + a$$

Terms are then rearranged to solve for **a** as the linear combination of the five known values, the two slopes and the three means:

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

$$a = 8 - 0.463\,(6) - 1.210\,(4)$$

$$= 8 - 2.779 - 4.840 = 8 - 7.619$$

$$= 0.381$$

The Y-intercept, Y', is therefore found to be 0.381.

A5.2.2.6 Write the Multiple Regression Predictive Equation

With the two slopes $b_1 = 0.473$ and $b_2 = 1.210$ and the Y-intercept of $\boldsymbol{a} = 0.381$, the multiple regression predictive equation can now be created for this data example.

$$\hat{\boldsymbol{Y}} = 0.463X_1 + 1.210X_2 + 0.381$$

A5.2.3 Step 3. Create Predicted Scores on Y from Each of the Seven Pairs of X_1 and X_2 Scores

This predictive equation can be used for predicting harassment, Y, for one starship at a time. We will test it with Starship D, the one that had the highest actual harassment score of $Y = 12$. It also had the highest competitive-climate predictor score of $X_1 = 11$, and a medium numerical dominance predictor score of $X_2 = 4$. The Stata analysis returned a predicted harassment score of 10.316 for Starship D, which agrees with the following calculations from the predictive equation.

$$\hat{\boldsymbol{Y}} = 0.463(11) + 1.210(4) + 0.381$$
$$= 5.095 + 4.840 + 0.381 = 10.316$$

Starship F had the lowest actual harassment score of $Y = 4$, a low competitive-climate predictor score of $X_1 = 4$, and a relatively low numerical dominance predictor score of $X_2 = 3$. Stata analysis returned a predicted harassment score of 5.864 for Starship F, which also agrees with the calculations from the predictive equation.

$$\hat{\boldsymbol{Y}} = 0.463(4) + 1.210(3) + 0.381$$
$$= 1.853 + 3.630 + 0.381 = 5.864$$

These calculations are more easily understood and more efficiently carried out using the matrix equation for the calculation of all predicted scores simultaneously, $\hat{\boldsymbol{Y}} = \boldsymbol{Xb} + \boldsymbol{a}$.

$$\underset{\hat{\boldsymbol{Y}}}{\begin{bmatrix} 11.167 \\ 8.104 \\ 6.147 \\ 10.316 \\ 8.179 \\ 5.864 \\ 6.222 \end{bmatrix}} = \underset{\boldsymbol{X}}{\begin{bmatrix} 5 & 7 \\ 1 & 6 \\ 2 & 4 \\ 11 & 4 \\ 9 & 3 \\ 4 & 3 \\ 10 & 1 \end{bmatrix}} \times \underset{\boldsymbol{b}}{\begin{bmatrix} 0.463 \\ 1.210 \end{bmatrix}} + \underset{\boldsymbol{a}}{\begin{bmatrix} 0.381 \\ 0.381 \\ 0.381 \\ 0.381 \\ 0.381 \\ 0.381 \\ 0.381 \end{bmatrix}}$$

The matrix calculations of predicted scores, $\hat{\boldsymbol{Y}}$, for all seven starships agree to three-place accuracy with the values from Figure A5.6, the screen capture of Stata's multiple regression data spreadsheet above, which are also shown graphically in Figure A5.5.

A5.2.4 Step 4. Calculate SSR, the Sum of Squares Regression

The deviation score form of the vector of predicted scores is created by centering, that is, by subtracting the vector containing the Y mean from the vector of raw predicted scores.

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}} = \begin{bmatrix} 11.167 \\ 8.104 \\ 6.147 \\ 10.316 \\ 8.179 \\ 5.864 \\ 6.222 \end{bmatrix} - \begin{bmatrix} 8 \\ 8 \\ 8 \\ 8 \\ 8 \\ 8 \\ 8 \end{bmatrix} = \begin{bmatrix} 3.167 \\ 0.104 \\ -1.853 \\ 2.316 \\ 0.179 \\ -2.136 \\ -1.778 \end{bmatrix}$$

The sum of squares regression, *SSR*, is calculated as the matrix product of the deviation scores vector, $\hat{\boldsymbol{y}}$, premultiplied by its own transpose $\hat{\boldsymbol{y}}'$.

$$SSR = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} = \begin{bmatrix} 3.167 & 0.104 & -1.853 & 2.316 & 0.179 & -2.136 & -1.778 \end{bmatrix} \begin{bmatrix} 3.167 \\ 0.104 \\ -1.853 \\ 2.316 \\ 0.179 \\ -2.136 \\ -1.778 \end{bmatrix} = 26.592$$

A5.2.5 Step 5. Calculate the Errors of Prediction by Subtracting Predicted Scores, $\hat{Y}$, from Actual *Y* Scores

The vector of the errors of prediction is calculated by subtracting the vector of predicted scores, $\hat{Y}$ from the vector of the actual **Y** scores, $e = Y - \hat{Y}$.

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \begin{bmatrix} 9 \\ 10 \\ 7 \\ 12 \\ 8 \\ 4 \\ 6 \end{bmatrix} - \begin{bmatrix} 11.167 \\ 8.104 \\ 6.147 \\ 10.316 \\ 8.179 \\ 5.864 \\ 6.222 \end{bmatrix} = \begin{bmatrix} -2.167 \\ 1.896 \\ 0.853 \\ 1.684 \\ -0.179 \\ -1.864 \\ -0.222 \end{bmatrix}$$

A5.2.6 Calculate a Simple Structure, that is, *a Clear and Interpretable Factor Solution*. However, SSE, the Sum of Squares Error

The sum of squares error, SSE, is calculated as the product of the 7×1 error scores vector, **e**, which is of course already in deviation score form, premultiplied by its own 1×7 transpose, $\mathbf{e}'$.

$$\boldsymbol{SSE} = \boldsymbol{e}'\boldsymbol{e} = \begin{bmatrix} -2.167 & 1.896 & 0.853 & 1.684 & -0.179 & -1.864 & -0.222 \end{bmatrix} \begin{bmatrix} -2.167 \\ 1.896 \\ 0.853 \\ 1.684 \\ -0.179 \\ -1.864 \\ -0.222 \end{bmatrix} = 15.408$$

A5.2.7 Step 7. Confirm that *SST* = *SSR* + *SSE*, and also that the Squared Multiple Correlation Is Equal to R^2 = *SSR/SST*, and that One Minus the Squared Multiple Correlation Is Equal to 1 − R^2 = *SSE/SST*

The formula for the squared multiple correlation coefficient is $R^2 = \boldsymbol{r}_{yx}\boldsymbol{R}_{xx}^{-1}\boldsymbol{r}_{xy}$. For this dataset, it is found to be $R^2 = 0.6331$, which agrees with the value reported as "R-squared" in the multiple regression Stata output table above.

$$R^2 = \boldsymbol{r}_{yx}\boldsymbol{R}_{xx}^{-1}\boldsymbol{r}_{xy} = \begin{bmatrix}0.2047 & 0.5357\end{bmatrix}\begin{bmatrix}1.4153 & 0.7667\\ 0.7667 & 1.4153\end{bmatrix}\begin{bmatrix}0.2047\\ 0.5357\end{bmatrix}$$

$$= \begin{bmatrix}0.7004 & 0.9151\end{bmatrix}\begin{bmatrix}0.2047\\ 0.5357\end{bmatrix} = 0.6331$$

This is analogous to the coefficient of determination in bivariate regression, and $1 - R^2$ is analogous to the coefficient of nondetermination. $1 - R^2 = 1 - 0.6331 = 0.3669$.

From the calculations up to this point, it can now be confirmed that the sum of squares total (SST) is equal to the SSR plus the SSE.

$$SST = SSR + SSE = 26.592 + 15.408 = 42.000$$

The squared multiple correlation, R^2, the coefficient of determination for multiple regression, can also be shown to be equal to the SSR divided by the SST. That is, the squared multiple correlation can be literally interpreted as the proportion of the total sum of squares (42.000) that is accounted for by the predicted scores in multiple regression (26.592).

$$R^2 = SSR/SST = 26.592/42.000 = 0.6331$$

One minus the squared multiple correlation, $1 - R^2$, the coefficient of nondetermination for multiple regression, is found to be equal to the SSE divided by the SST, indicating that it can be literally interpreted as the proportion of the total sum squares due to error, that is, the proportion that is not accounted for by multiple regression.

$$1 - R^2 = \text{SSE/SST} = 15.408/42.000 = 0.3669$$

A5.2.8 Step 8. Calculate the ANOVA Table Statistics: The Three Mean Squares (MSR, MSE, and MST), the *F* Ratio, and the Standard Error of the Estimate, $s_{y.x}$

The mean square regression, MSR, is found by dividing the SSR by its degrees of freedom, $dfr = h - 1$, where h is the number of independent parameters that are used to create the regression line. In this two-predictor multiple regression model h is 3, the two slopes and the intercept. The degrees of freedom is $h - 1$ because the predicted scores, like the observed scores, are constrained by their mean. In other words, the sum of the predicted scores around their mean is zero.

$$MSR = \frac{SSR}{h-1} = \frac{26.592}{3-1} = \frac{26.592}{2} = 13.296$$

The degrees of freedom for the error sum of squares, SSE, is $dfe = N - h$, because the deviation of the N observed scores from the regression line is constrained by the three parameters used to create the regression line in this two-predictor example of multiple regression. The mean square error, MSE, is found by dividing SSE by its degrees of freedom, $dfe = N - h$.

$$MSE = \frac{SSE}{N-h} = \frac{15.408}{7-3} = \frac{15.408}{4} = 3.852$$

The total degrees of freedom, *dft*, is, of course, $N - 1$. There are N observations but only $N - 1$ of them are free to vary since the deviations from the mean must sum to zero. The mean square total, MST, is found by dividing SST by its degrees of freedom.

$$MST = \frac{SST}{N-1} = \frac{42.000}{7-1} = \frac{42.000}{6} = 7.000$$

An F ratio can be calculated by dividing MSR by MSE to test this multiple regression model of the starship dataset for statistical significance.

$$F = \frac{SSR}{SSE} = \frac{13.296}{3.852} = 3.45$$

This F ratio has 2 degrees of freedom in the numerator and 4 degrees of freedom in the denominator. So, the 0.05 critical ratio is $F\,(0.05, 2, 4) = 6.94$. The obtained F ratio does not exceed the critical ratio. The null hypothesis is therefore not rejected.

The standard error of the estimate for this multiple regression model, s_e, is found by taking the square root of MSE.

$$s_e = \sqrt{MSE} = \sqrt{3.852} = 1.9626$$

This agrees with the standard error in the Stata output table, as given in the middle of Part 1 of this appendix. It is found in the upper right corner of the table, below the F ratio, R^2, and other summary statistics, with the label "Root MSE," which stands for the square root of the MSE. Part 3 will consider this same multiple regression analysis but with a larger version of the starship dataset.

A5.3 Multiple Regression Results with a Reasonably Larger Sample

In the foregoing multiple regression analysis, a very small sample size was used to simplify the calculations, thus making the concepts of multiple regression and also the calculations substantially easier to understand. The downside of this is that it might suggest that regression can be run on very small datasets. Many researchers say that a good rule of thumb is that there should be at least 10 observations for each predictor variable in the model. With this particular design, there were two predictor variables. To have more stable results, it would be good to have 20 or more starships in the story problem. Looking at the foregoing results with a sample size of 7, it can be seen that even though a reasonably strong R^2 value of 0.6331 is obtained, with only 7 starships, the F ratio is only 3.45 and does not reach statistical significance.

If the sample size were to be twice as large, with an N of 14, a statistically significant result could be expected with this same pattern of data. To test that conjecture, Stata code is used to create a simulated dataset with the same data pattern, but with a sample size of 14. The following six lines of Stata code will create an $N = 14$ dataset having the same specified correlation matrix, means, and standard deviations as the $N = 7$ dataset.

```
clear
matrix C = (1,.2047,.5354\.2047,1,-.5416\.5354,-.5416,1)
corr2data y x1 x2, n(14) corr(C) means(8 6 4) sds(2.648 4 2)
list
summarize
corr
corr, covariance
```

Running the first three of these lines of Stata code will create the simulated dataset. The first line specifies the 3×3 correlation MATRIX of the new dataset. The second line specifies the labels of the three variables (y, $x1$, $x2$), the sample size ($n(14)$), the means, and the standard deviations (sds). Once these three lines of code are run a new simulated dataset appears in the Stata data editor and in the Stata output, similar to the one shown below. If this same Stata code were to be run again, the obtained data values would not be exactly the same. The pattern changes in its particulars every time this Stata "data reconstruction" code is run, but the summary statistics and the regression output statistics will be the same.

	y	*x*1	*x*2
1	3.6241	8.0691	2.6100
2	8.3778	1.2058	5.4352
3	12.5993	11.7618	4.1874
4	7.5325	–0.7618	7.4988
5	6.6440	11.4199	1.6155
6	9.8128	5.5974	7.0780
7	6.8263	0.2948	3.7121
8	4.3460	4.8238	1.1946
9	6.6021	4.6302	3.7368
10	9.5117	7.1572	3.9823
11	11.5017	10.0070	3.8079
12	5.2709	9.6933	0.8717
13	10.4082	6.5225	4.8617
14	8.9424	3.4962	5.4081

The fourth command, "summarize," will print the following summary statistics for the newly created dataset, including variables, number of observations, means, standard deviations, minimum scores, and maximum scores.

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	14	8	2.648	3.6241	12.59932
x1	14	6	4	-.6791058	11.76184
x2	14	4	2	.8716609	7.4988

The fifth command, "corr," will print the correlation matrix.

```
. corr
(obs=21)
```

	y	x1	x2
y	1.0000		
x1	0.2047	1.0000	
x2	0.5354	-.5416	1.0000

The sixth command, "corr, covariance" will print the covariance matrix.

```
. corr, covariance
(obs=21)
```

	y	x1	x2
y	7.0119		
x1	2.16818	16	
x2	2.83548	-4.3328	1.0000

The data will remain in the Stata data editor until it is closed out by quitting Stata or by entering the Stata command "clear." With this new $N = 14$ dataset still in Stata, entering the same command originally entered to calculate the multiple regression on the $N = 7$ starship data will create the multiple regression output for an $N = 14$ version of the same data structure, with this result.

regress y x1 x2

```
. regress y x1 x2
```

Source	SS	df	MS
Model	57.6942963	2	28.8471481
Residual	33.4604598	11	3.04185998
Total	91.1547561	13	7.01190431

Number of obs	= 14
F(2, 11)	= 9.48
Prob > F	= 0.0040
R-squared	= 0.6392
Adj R-squared	= 0.5662
Root MSE	= 1.7441

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	.4634038	.1438566	3.22	0.008	.1467776	.7800301
x2	1.210829	.2877132	4.21	0.001	.5775762	1.844081
_cons	.3762622	1.833778	0.21	0.841	-3.659856	4.41238

Notice that the R^2 value in this $N = 14$ dataset is about the same as it was in the $N = 7$ dataset output table in the middle of Part 1 (0.6329 compared to 0.6331). However, all of the other statistics have changed considerably, with a substantially larger F ratio (9.48 compared to 3.45), and a statistically significant value (0.0040 compared to 0.1346). Even the adjusted R^2 increases from 0.4497 to 0.5662. The t values for the two predictors are both now statistically significant ($t = 3.22$, $p = 0.008$, and $t = 4.21$, $p = 0.001$) as are, of course, the two corresponding confidence intervals. The two slopes ("Coeff") for the two predictors of course remain about the same, but the standard errors of the two slopes are substantially smaller (0.1439 compared to 0.2383, and 0.2877 compared to 0.4766).

A5.4 Discussion

Multiple regression was one of the major developments of the 20th century in statistical science. It has now become one of the main "work horses" of statistical methods. It is ubiquitous. It has come a long way since its earliest beginnings in the last decade of the 18th century when Carl Friedrich Gauss first used the least squares method to model the path of a comet as it passed behind the sun. A century later, Karl Pearson's creation of the product moment correlation coefficient

represented the second major step in the development of multiple regression method as it exists today. Another major contribution to multiple regression, as we know it today, was the Cholesky decomposition method for fitting simultaneous equations, published posthumously after its author, André-Louis Cholesky of the French army, died in battle during WW I. It has become central to the functioning of many machine learning and Monte Carlo simulation models, but its major usefulness is in its centrality to multiple regression models. It is a particularly efficient method for creating the inverse of a matrix, the computational engine for multiple regression. In section 29 of Chapter 3, it is proposed that in addition to its superior practical usefulness as an empirical tool, multiple regression also serves as an illuminating statistical metaphor for scientific explanation and prediction.

A.5.5 Study Questions for Appendix 5

Multiple Regression

Essay Questions

1. Summarize the geometry of the "graphical basis of multiple regression" and compare it to the geometry of the "graphical basis of bivariate regression." You may wish to include drawings.
2. How does multiple regression build upon and go beyond bivariate regression mathematically?
3. What additional matrix algebra calculations does it employ?
4. In what other ways are the calculation steps expanded?
5. How does multiple regression go beyond bivariate regression in the kinds of questions it can answer?
6. How do the predictive equations for multiple regression and bivariate regression compare?
7. How does the standard error of the estimate for multiple regression compare to the standard error of the estimate for bivariate regression?
8. How does multiple regression compare to multiple correlation?
9. How does the coefficient of determination for a multiple regression with two predictors compare to the coefficient of determination for a bivariate regression that only employs one of those two predictors?
10. How does the coefficient of nondetermination for a multiple regression with two predictor variables compare to the coefficient of nondetermination for a bivariate regression that only employs one of those two predictors?
11. If a third predictor variable were added to the multiple regression with two predictors, how would that effect the coefficient of determination? How would it affect the coefficient of nondetermination?
12. How do the scatterplots for multiple regression compare to those for bivariate regression?

Calculation Questions

1. From start to finish, calculate the entire multiple regression analysis demonstrated in this appendix, predicting the number of harassment incidents from the two predictive variables, "team climate" and "team numerical dominance" using a spreadsheet such as excel. Begin with the small dataset of $N = 7$ (Table A5.1) from which the multiple regression demonstration in this appendix was constructed.

2 Now use Stata to run a multiple regression analysis on this same dataset. Compare your results to those obtained from your analysis using the spreadsheet and also compare to the results given in the appendix. Do the results agree?

3 In part 3 of the appendix titled "Multiple Regression Results with a Larger Sample," you were shown how to use Stata to create a simulation based upon the summary statistics (means, standard deviations, and correlation coefficients) of a small sample to create a larger sample having the same summary statistics as your small sample. Create comparable Stata code and use it to transform your $N = 7$ dataset into a dataset twice as large, $N = 14$. Compare your new $N = 14$ dataset with the one shown in this appendix. Because this is a stochastic simulation (a simulation based upon the laws of probability), each dataset created using this Stata code will differ in particulars, yet each will give the same results to many decimal places when analyzed with the multiple regression Stata code instructions. The various runs done in this same way with the same sample size of $N = 14$ will be "replications" of one another.

4 Now run the multiple regression on this new $N = 14$ dataset using the Stata command **regress y x1 x2**. You will notice that your "multiple regression level" results will be compatible to many decimal places with those shown in the $N = 14$ results from part 3 of the appendix even though the datasets will differ from one another in their particulars, that is, their individual values.

5 The CERAD neurological battery is a multidimensional instrument created to diagnose, evaluate, and monitor Alzheimer's disease.[4] We will use it here as a second practice example for running and interpreting a multiple regression. The following Stata code can be used, following the procedure outlined in part 3 of this appendix, to create a small dataset with 6 respondents and four psychological tests. This is just like what you did in question 3 above, except that you will now have one less respondent and one more predictor variable. You will be predicting a single y variable from $x1$, $x2$, and $x3$. Enter this "do-file" into the "do-file editor," and run it. You will see the 6×4 matrix (6 rows and 4 columns) of your dataset appear on the Stata screen.

```
clear
matrix C = (1,.50,.52,.43\.50,1,.40,.54\.52,.40,1,.50\.43,.54,.50,1)
corr2data y x1 x2 x3, n(14) corr(C) means(11.1 11.9 23.6 4.3) sds
(5.4 3.0 11.8 2.0)
list
```

6 Now enter the command "**summarize**" into the command line at the bottom, below the screen, and hit the "return" button, and you will see a table with "Variable," "Obs" (number of observations), Mean, Std. dev (standard deviations), Min, and Max (the minimum and the maximum values in each of the four variables).

7 Now enter the command "**corr**" in the same way and you will see the 4×4 correlation matrix among these four variables.

8 Now enter the command "**corr, covariance**" in the same way and you will see the 4×4 covariance matrix among these four variables.

4 The Consortium to Establish a Registry for Alzheimer's disease (CERAD) was established in 1986 by a grant from the National Institute on Aging (NIA) to standardize measurement procedures in this area. Eight tables in Chapter 7 (tables 7.2–7.9) illustrate various factoring methods applied to analysis of scores on tests from this battery. This calculation question provides practice in using multiple regression but also provides a backdrop comparison, when you read Chapter 7 for seeing how factoring goes beyond multiple regression.

9 Finally, complete the process by entering the command "**regress y x1 x2, x3**" and it will print the multiple regression analysis results for this dataset, comparable to the $N = 7$ small dataset that you ran in question 2 above and the large $N = 14$ dataset that you ran in question 4 above. On the basis of what you have learned from this appendix what differences and similarities do you notice among these three multiple regression analyses?

10 Now calculate the multiple regression analysis for this small dataset using Excel as you did for the demonstration dataset in question 1 above.

11 The four variables in the multiple regression you just ran on the CERAD data are y = "Verbal Fluency," x1 = "Boston Naming," x2 = "Control Oral Word," and x3 = "Word List." If you go to table 7.1 in chapter 7, you will see that the first four CERAD "subtests" listed are the same as the four variables and their correlations that you entered in question 5 above. These four variables constitute the "Verbal" cluster of the CERAD test shown in (a), in the left half of figure 7.4. These four subtests are in the same cluster because they are quite highly correlated with one another, with correlation coefficients ranging from 0.40 to 0.54. You will notice, however, from the multiple regression results of the question 9 analysis that although the multiple regression has an R^2 value of 0.3763, it has a very low F ratio of 0.40 and is not statistically significant. That is because it has a very small sample size of $N = 6$, which makes calculations easier to follow, but does not produce realistic results. We will now try it again with a somewhat more reasonable sample of $N = 30$, still on the smallish side, but more informative.

12 The following Stata "do-file" code, when entered into the "do-file editor" as we did in question 5 will create the $N = 30$ dataset with five times as many observations as we created in question 5. However, this larger dataset will have the same descriptive statistics (correlation coefficients, standard deviations, means, etc.) as found in the $N = 6$ dataset.

```
clear
matrix C = (1,.50,.52,.43\.50,1,.40,.54\.52,.40,1,.50\.43,.54,.50,1)
corr2data y x1 x2 x3, n(30) corr(C) means(11.1 11.9 23.6 4.3)
sds (5.4 3.0 11.8 2.0)
list
```

13 Now enter the command "**summarize**" into the Stata command line to see the summary statistics, then the commands "**corr**" and "**corr, covariance**" to view the 4×4 correlation matrix and the 4×4 covariance matrix. The only thing that differs is the second column from the output of the "**summarize**" command, "Obs," where we see that each variable has 30 observations rather than 6.

14 We now finalize this set of questions by running the multiple regression with the command "**regress yvf x1bn x2ow x3wl**."[5] Compare the multiple regression output tables for the $N = 6$ analysis and this $N = 30$ analysis. Obviously, the $N = 30$ analysis is a somewhat more realistic simulation, with the overall analysis reaching statistical significance at the 0.01 level, with the F ratio now at 5.23 and a probability of 0.0059.

5 Notice that we have changed the variable names in the simulation of this larger dataset to better identify the subtests of the CERAD on each of the printouts of correlation matrices, summary statistics, etc., *y* becomes *yvf*, where *vf* is the abbreviation for verbal fluency; *x*1 becomes *x1bn*, where *bn* is the abbreviation for Boston Naming; *x*2 becomes *x2ow*, where ow is the abbreviation for (control) oral word; and *x*3 becomes *x3wl*, where *wl* is the abbreviation for word list.

```
. regress Yvf x1bn x2ow x3wl
```

Source	SS	df	MS	Number of obs	=	30
				F(2, 26)	=	5.23
Model	318.204196	3	106.068065	Prob > F	=	0.0059
Residual	527.435789	26	20.2859919	R-squared	=	0.3763
				Adj R-squared	=	0.3043
Total	845.639985	29	29.1599995	Root MSE	=	4.504

Yvf	Coefficient	Std. err.	t	P>|t|	[95% Conf.	Interval]
x1bn	.5640746	.336634	1.68	0.106	-.1278865	1.256036
x2ow	.1612442	.0831775	1.94	0.063	-.0097296	.332218
x3wl	.2284292	.53439	0.43	0.673	-.8700251	1.326884
_cons	-.4000962	3.441629	-0.12	0.908	-7.474466	6.674273

15 A larger simulation would be with an *N* of 90, using the following Stata "do-file" code:

```
clear
matrix C = (1,.50,.52,.43\.50,1,.40,.54\.52,.40,1,.50\.43,.54,.50,1)
corr2data y x1 x2 x3, n(90) corr(C) means(11.1 11.9 23.6 4.3)
sds (5.4 3.0 11.8 2.0)
list
```

```
. regress Yvf x1bn x2ow x3wl
```

Source	SS	df	MS	Number of obs	=	90
				F(3, 86)	=	17.29
Model	976.557714	3	325.519238	Prob > F	=	0.0000
Residual	1618.6823	86	18.8218873	R-squared	=	0.3763
				Adj R-squared	=	0.3545
Total	2595.24002	89	29.1600002	Root MSE	=	4.3384

Yvf	Coefficient	Std. err.	t	P>|t|	[95% Conf.	Interval]
x1bn	.5640746	.1850953	3.05	0.003	.1961173	.932032
x2ow	.1612442	.0457344	3.53	0.001	.0703271	.2521612
x3wl	.2284293	.2938297	0.78	0.439	-.3556849	.8125434
_cons	-.4000962	1.893591	-0.21	0.833	-4.164431	3.364239

References for the Five Mathematical Appendices

Bravais, A. (1846). Analyse mathématique sur les probabilités des erreurs de situation d'un point. *Mémoires présentés par divers savants à l'Académie royale des sciences de l'Institut de France*, 9, 255–332.

Brown, B. L., Hendrix, S. B., Hedges, D. W., & Smith, T. B. (2012). *Multivariate Analysis for the Biobehavioral and Social Sciences: A Graphical Approach*. Hoboken: John Wiley & Sons.

Cholesky, A.-L. (2023). The Cholesky decomposition. When it is applicable, it is about twice as efficient as the LU decomposition for solving linear equation systems. See *Cholesky decomposition* in Wikipedia https://en.wikipedia.org/wiki/Cholesky_decomposition.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society*, 45, 135–145.

Gauss, C. F. (1794/1795). Footnote 1 in the Brown, Hendrix, Hedges, & Smith 2012 multivariate analysis book briefly outlines and refers to several sources detailing the priority dispute between Gauss and Legendre regarding the discovery of the method of least squares.

Horst, P. (1966). *Psychological Measurement and Prediction*, Belmont: Wadsworth Publishing Company

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society*. 187-A, 253–318.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13*(1),* 25–45.

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

Index

Note: Page numbers in *italics* refer to figures and *t* refer to tables.

a

accuracy 8, 23, 38, 145, 169, 218, 324, 327, 354
adjacent category models
  with Mplus
    GPCM 429–430
    GRSM 431
    model D 431–432
    PCM 430
  using Stata 424–425
Alzheimer's disease (AD) 70–71, 86–87, 133, 241
analysis of variance (ANOVA) 125, 378, 528–529, 552–553
anthropometric measurements 7
applied metrology 327–328
attributes 10
  "determining" *IQ* in children 129–130
  local uncorrelatedness and falsifiability 130–131, *131*
  sampling stratification 130
  statistical control 130
  statistical metaphor 129
average relative parameter bias 477–478
average variance extracted (AVE) 371

b

Bartlett's WLS method 244, 245*t*
behavioral sciences
  measurability in 23
  statistical methods 3–4
  variables (*see* variables)
behavioral statistics 4–5
Bell curve. *see* normal distribution
biased variance 498–499
bifactor models 278, *279*
  CFA 452–455, 454*t*, 455*t*
  EFA/ESEM 455, 456*t*
  vs. higher order model
    constraints 457–458
    extended higher order model 461–462, *462*
    loadings of indicators 458–461, 460*t*
  model fit of multidimensional factor models 456, 457*t*
  unidimensional and multidimensional factor analysis *442,* 442–444, *443*
binomial distribution
  calculations for 49–50, *50*
  independent joint events, multiplication rule *42,* 42–44, *44*
  multinomial distribution probability model 52–56, *53, 55*
  mutually exclusive composite events, addition rule 44–46
  sampling distribution of proportions 46–49, *47, 49*
Birnbaum, Allan 78, 79
bivariate regression analysis
  ANOVA table statistics 528–529
  correlation coefficient 521–523
  covariance matrix 521–523
  errors-of-prediction scores 526–527
  predicted scores on Y 524–525, *525*
  review questions 531–533

Psychometrics, Test Theory, and the Latent Factors Model, First Edition. Petr Blahuš, Bruce L. Brown, C. Victor Bunderson, and Joseph A. Olsen.
© 2026 John Wiley & Sons, Inc. All rights reserved, including rights for text and data mining and training of artificial intelligence technologies or similar technologies. Published 2026 by John Wiley & Sons, Inc.
Companion website: www.wiley.com/go/psychometrics

bivariate regression analysis (*cont'd*)
scatterplots *520,* 520–521
slope and *Y* intercept 523–524
SSE 527–528
SSR 526–528
SST 527–528
standard deviation 521–523
Stata 529–531, 529*t*, *530, 531*
Borsboom, D. 104

C

calibration, measurement error 162, 163, 170
categorical item factor analysis (CIFA) 406–407
dichotomous items for *408*
1PL and 2PL models 412, 412*t*
ICCs and IIFs 412–413, *413*
parameterization 408–410, 409t
Rasch model 410–412, 411*t*
ordered polytomous categorical items 418, 419*t*
causal variables *137,* 137–139, *138*
causation
causal explanations 196–197
declarative knowledge 195
observed performance ratings 195
repeated trials and coaching 197
socio-cognitive theory 197
target domain 196
centered cross product (CCP) 521
centered sum of squares (CSS) 521–522
central tendency 485–489
mean 485–489
median 485–489
mode 485–489
centroid method 234–238, 235*t*–237*t*
CFA. *see* confirmatory factor analysis (CFA)
chi-square difference test 363, *364,* 364*t*
CIFA. *see* categorical item factor analysis (CIFA)
classical test theory (CTT) 5, 24–25, 247
application of 176–177, 177*t,* 181–182
axiom of classical parallelism 179–180
axiom of counter-balanced errors 179
axiom of uncorrelated errors 179
axiom of uncorrelatedness of errors with true values 179
covariance 180–181
expectation principle 177–178
formulas 180
observed scores 180
parallel forms 176
problems in 182
procedures 175
randomness principle 178–179
reliability (*see* reliability)
true score models (*see* true score models)
true scores and errors 180
coefficient Alpha
bootstrapping 376
congeneric data 377–378
estimation of 378–379
Mplus output 376
parameters 375
statistical modeling program 375
coefficient Omega
vs. Alpha 380–382
bootstrapping 380
item deletion 382–383
Mplus syntax 379, 380
coefficient Rho 383–384
coefficients of determination 244
coefficients of nondetermination 244
column vector 506, 514
communality
CTT 247
SMC 249–250
specificity and unreliability 248–249
comparative fit index (CFI) 286
computational guides 350–351
computer-adaptive testing
high-jumping metaphor 206–207
item characteristic curves 204, *205*
Wright Map 204–206, *205*
confirmatory approach 287–288, 289*t,* 308. *see also* exploratory approach
confirmatory factor analysis (CFA) 286, 313
bifactor models 452–455, 454*t,* 455*t*
correlated factors model 448, 448*t*
higher-order factor models 448–451, 449*t,* 450*t*
unidimensional and multidimensional factor analysis 439, 440, *440*
congeneric model, true score models
equally reliable indicators 372–374, 373*t,* 374*t*
Mplus syntax 368–369

Consortium to Establish a Registry for AD (CERAD) 70–71
Constructional Praxis (C-Praxis) 70–71
construction formula 227
construct replicability index 477
construct validity. *see also* internal validity
  content and 198–200
  latent variables 132–133, 133*t*
  LCF model 134
  local uncorrelatedness 132–133, 133*t*
content homogeneity 241, 308–309
  equivocal indicators and factorial complexity 296
  higher-order factor analysis 295–296, 297*t*
  stages 294–295
"content-homogeneous" coordination tests 226
content validity
  domain theory 192–195
  item response theory
    causation 195–197
    computer-adaptive testing 204–207, *205*
    indicators/subtests 198–204
  latent factor 192–195
  latent variable 207
    class models 215–216
    common factor 214
    first- and second-order moments 216–217
    full-information method 217
    homogeneous indicators 208–211
    latent ability 214–215
    level of quantification 212–214
    local probability independence 217
    mathematical formulation 216
    normal distribution 216
    objects modeling 211–212
    partial-information method 217
    trait models 214
    uncorrelatedness 217
  review questions 220–221
correlated factors model
  CFA 448, 448*t*
  communalities and ESEM in Mplus 446–447, 448*t*
  EFA 445–446, 446*t,* 447*t*
  reliability index for *464,* 464–466, *465*
correlated indicator residuals 481, 481*t*–482*t*
correlation
  vs. causation 125
    counterfactual condition 127
    experimental manipulation 126–127
    passive observation 126–127
  coefficient
    bivariate regression 521–523
    historical development 5–6
    with regression and heritability 7–9
    review questions 57
  composite validity 115–118
  matrix, multiple regression 545–546
COSAN 350
counterfactual condition 127
covariance 500–501
covariance matrix, bivariate regression 521–523
Cronbach, Lee J. 102
CTT. *see* classical test theory (CTT)
cumulative density function
  computation for 78–80, *79*
  need for 77–78
cumulative probability models 432
  using Mplus
    GRM 426–427
    GRM-C 427
    MGRM 427–428
    MGRM-C 428–429
  using Stata 423–424

d

data variables 9–11
development programs 158
dimensionality 280
  factor rotation and simple structure 284–285
  ICS 282–283, *283*
  LCF *vs.* PCA 309
    "communalities" in PCA-2 303, 304*t*
    correlations *vs.* variance-1 300, 301*t,* 302*t,* 303
    deterministic component scores *vs.* factor score indeterminacy 305–306
    "variance" in PCA-3 303, 305
Direct Measurement 229, 230
discrepancy, LCFs
  least squares fitting and model fit assessment 239–240
  maximum likelihood 242, *243*

discrepancy, LCFs (*cont'd*)
relative model fit 240–241
unidimensionality and content homogeneity 241
vague formulation 239
WLS fitting 241–242
discriminability
standard deviation 66–68
standardized scores 87
distribution of scores
quantification (*see* quantification)
review questions 93–96
standard scores 82–91
domain-specific analysis and modeling 351–352
domain theory
domain-specific analysis and modeling 351–352
educational and psychological properties 156
FORE domain 193–194
goal of 155–156
internal aspects of validity 193
learning progression 194
model-based reasoning 194–195
multiple revisions 193
ostensive demonstration of concatenation 352–353
physical properties 156–157
skills and abilities 156
uses 157
Domotor, Zoltan 17–18

e

ECD. *see* evidence-centered design (ECD)
ECV. *see* expected common variance (ECV)
educational evaluation 355–357, 356*t*
EFA. *see* exploratory factor analysis (EFA)
effect variables 137, *137*
empirical sciences, mathematical methods 3
errors of observation 6
essentially parallel model 361, 363, 370
European Accreditation of Laboratories (EAL) 171
evidence-centered design (ECD) 106, 108
design for 159
documentation and resources for 160
learning progression 194
scientific standards 158–159
traditional trade practices 158
expanded uncertainty 171
expected common variance (ECV)
for general factor 475
item 475–477
for subscales 475
expected value 491–493, 492*t,* 493*t*
exploratory approach
factor analysis 291–292
factor rotation 292, 293*t*, *294*
simple structure 288–291, 290*t,* 308
exploratory factor analysis (EFA) 283, 312
bifactor models 455, 456*t*
correlated factors model 445–446, 446*t,* 447*t*
higher-order factor models 451–452, 451*t*–452*t*
unidimensional and multidimensional factor analysis 439, *440,* 441, *441*
exploratory structural equation modeling (ESEM) approach 446–447, 448*t,* 451–452, 451*t*–452*t,* 455, 456*t*
extensive metadomains 332–333
external validity
aspects 112
composite validity 115
correlation 127–128, 128*t*
criterion-related validity 113–114
maximal validity 116–117
multiple correlation 117–118
multiple regression 118–121
overview 111–112
standard error of estimation 114
standardized composite scores 112–113
subtests 116
variability and 115
weighted composite validity 116

f

factor analysis 2
bifactor models
CFA 452–455, 454*t,* 455*t*
EFA/ESEM 455, 456*t*
vs. higher order model 457–462
model fit of multidimensional factor models 456, 457*t*
unidimensional and multidimensional factor analysis *442,* 442–444, *443*

correlated factors model
CFA 448, 448*t*
communalities and ESEM in Mplus 446–447, 448*t*
EFA 445–446, 446*t,* 447*t*
reliability index for *464,* 464–466, *465*
higher-order factor models
vs. bifactor models 457–462
CFA 448–451, 449*t,* 450*t*
EFA/ESEM model 451–452, 451*t*–452*t*
unidimensional and multidimensional factor analysis *441,* 441–442, *442*
Omega reliability 472–475
reliability estimation, multidimensional constructs
higher-order models, analytic formulas 469–470
reliability index 467–469
stratified alpha 462–464, *464*
unidimensional and multidimensional factor analysis
bifactor models *442,* 442–444, *443*
CFA 439, 440, *440*
data example 444–445
EFA 439, *440,* 441, *441*
higher-order factor models *441,* 441–442, *442*
with multiple factors 438–439
factor determinacy 478–480
factor indeterminacy 243
factoring 2
Ferguson Committee 320–321
Fisher's information and efficiency 173–174
Fluent Oral Reading (FORE) 229–230
Fosbury Flop 348–349
frame problem 299–300
Fundamental Factor Theorem 238–239

g

Galton, Francis 5, 6
Gauss model for replications 164–178, 184
generalized partial credit model (GPCM) 429–430
generalized rating scale model (GRSM) 425, 431
generic reliability 253, 255–256
Goodness of Fit Indices (GFI) 240–241
graded response model (GRM) 423–424, 426–427
graded response model with common discrimination (GRM-C) 424, 427

h

heritability, correlation and regression 7–9
hidden equivocality 309
frame problem 299–300
indicator multidimensionality 298–299
overview 296, 298
hierarchical models 278–280, *279*
higher-order constructs 280
higher-order factor analysis 295–296, 297*t*
higher-order factor models
vs. bifactor models
constraints 457–458
extended higher order model 461–462, *462*
loadings of indicators 458–461, 460*t*
CFA 448–451, 449*t,* 450*t*
EFA/ESEM model 451–452, 451*t*–452*t*
unidimensional and multidimensional factor analysis *441,* 441–442, *442*
high jumping 348–349
implementation examples 207
measurement instrument 206
metric/standardized units 207
skill of 206
Holzinger–Swineford data 453, 454, 476
homogeneity
domain-specific model 208–209
factor rotation and simple structure 284–285
formula score 210
ICS 282–283, *283*
item analysis and factor analysis 211
performance 209
test blueprints 209–210
vs. unidimensionality 201–202
hypothetico-deductive approach 287

i

ICCs. *see* item characteristic curves (ICCs)
IIFs. *see* item information functions (IIFs)
imprecision 169–170
independent cluster structure (ICS) 282–283, *283*
equivocal indicators and factorial complexity 296
hidden equivocality 296, 298
higher-order factor analysis 295–296, 297*t*

instrument indication 162
instrument scale application 145
intensive metadomains 333
internal manifest validity 285
internal validity
  causal effects *vs.* independent/dependent variables 136–139, *137, 138*
  causal interpretation 139–140
  correlation 127–128, 128*t*
  empirical and observable 134–135
  formalized and unformalized levels 147
  latent variables 132–133, 133*t,* 135–136
  LCF model 134
  local uncorrelatedness 132–133, 133*t*
  manifest-latent link 142–146, 143*t*
  measurement approach 141–142
  partial-local uncorrelatedness 135–136
  propensity toward behavior *140,* 140–141
  theoretical and unobservable 135
International Bureau of Weights and Measures (BIPM) 323, 327
International Organization for Legal Metrology (OIML) 328
IRT. *see* item response theory (IRT)
item analysis 211, 214
item characteristic curves (ICCs) 204, *205,* 412–413, *413*
item information functions (IIFs) 412–413, *413*
item reliability 371–372
item response theory (IRT) 20, 193, 227, 253
  adjacent category models using Stata 424–425, 434*t*
  adjacent category models with Mplus
    GPCM 429–430
    GRSM 431
    model D 431–432
    PCM 430
  causation 195–197
  CIFA (*see* categorical item factor analysis (CIFA))
  computer-adaptive testing
    high-jumping metaphor 206–207
    item characteristic curves 204, *205*
    Wright Map 204–206, *205*
  cumulative probability models using Mplus
    GRM 426–427
    GRM-C 427
    MGRM 427–428
    MGRM-C 428–429
  cumulative probability models using Stata 434*t*
    GRM 423–424
    GRM-C 424
  data
    CIFA 406–407
    compact data and analyzing expanded data 404–405
    compressed data analysis 403–404
    Rasch model 405–406
  dichotomous items 400, *401*
  indicators/subtests
    agreement of experts 200–201
    content and construct validity 198–200
    content homogeneity *vs.* unidimensionality 201–202
    linear factor analysis 202–204
  logistic (logit) models
    1PL model 401
    2PL model 402, *402*
    Rasch model 401
  model implementation
    adjacent category models 432–433
    cumulative probability models 432
  normal ogive (probit) IRT models *414,* 414–415
  ordered polytomous items
    category comparison methods 417–418, 418*t*
    CIFA 418, 419*t*
    four ordinal models 421–422
    model equivalence, nesting, and comparisons 419–421, *420,* 420*t,* 421*t*
    model specification 418, 419*t*
  software and example analyses 422–423
  using Mplus
    1PL model 416–417
    2PL model 417
    Rasch model 415, *416,* 417

k

Kahneman, D. 85
Kane, Michael 103–108
Kinanthropometry 100
Krantz, D. M. 2, 16, 56

l

latent common factors (LCFs) model 192
aspects 134
communality
CTT 247
SMC 249–250
specificity and unreliability 248–249
discrepancy
least squares fitting and model fit assessment 239–240
maximum likelihood 242, *243*
relative model fit 240–241
unidimensionality and content homogeneity 241
vague formulation 239
WLS fitting 241–242
factor scores
estimation 244, 245*t,* 246
factor indeterminacy 243
generic true scores 247
latent values 246
FORE domain 193–194
generic reliability and "model–remodel" assessment 253, 255–256
internal aspects of validity 193
learning progression 194
loadings and residual correlations
centroid method 234–238, 235*t*–237*t*
computational optimization 234
Fundamental Factor Theorem 238–239
PC model and ML model 235, 236*t*–237*t,* 238
verbal tests 232, 233*t*
measurement and correspondence rules
analogs 226, *227*
overview 226
semantic correspondence rules 226–227
syntactic correspondence rules 227–230, *228*
model-based reasoning 194–195
multiple revisions 193
vs. PCA
"communalities" in PCA-2 303, 304t
correlations *vs.* variance-1, 300, 301*t,* 302*t,* 303
deterministic component scores *vs.* factor score indeterminacy 305–306
"variance" in PCA-3 303, 305
prototype of 230–232
quantitative exercise questions 258–260
regression estimate of *f* and decomposition of indicator variance
tapping test 251, 252*t*
test–retest analysis 251, 253, 253*t,* 254*t*, *255*
SPSS commands 260–269
Stata commands 269–274
latent trait theory 20
latent variable, factor analysis & uncorrelatedness 129, 132–136, 147
latent variable modeling 207
class models 215–216
common factor 214
first- and second-order moments 216–217
full-information method 217
homogeneous indicators 208–211
latent ability 214–215
level of quantification 212–214
local uncorrelatedness and 132–133, 133*t*
manifest variables and 142–144, 143*t*
mathematical formulation 216
normal distribution 216
objects modeling 211–212
partial-information method 217
principle of partial-local uncorrelatedness 135–136
theoretical concept formation 135
trait models 214
uncorrelatedness local probability independence 217
LCFs. *see* latent common factors (LCFs) model
least squares fitting 239–240
least squares regression, historical development 5–6
legal metrology 327–328
linear transformation, standardized scores 81
local uncorrelatedness 132–133, 135–136
logistic cumulative distribution function 79–80
lognormal distribution 87
Luce, R. D. 2, 16, 56

m

MAD. *see* mean absolute deviation (MAD)
marathon runners *84,* 84–85, *85*
Mari, L. 108–110, 141, 144, 146
Markus, K. A. 104

mathematical methods, empirical sciences 3
matrix algebra 349
matrix multiplication
  calculation for variance 509–511
  Excel spreadsheet 508–509
  history 505
  matrices, vectors, and scalars 506–508
  probabilities-weighting method 513–516, 514*t*, *516*
  review questions 516–518
  3 × 3 covariance matrix 511–513
Maul, A. 108–110, 141, 144, 146
maximum likelihood (ML) 235, 236*t*–237*t,* 238, 242, *243*
mean 485–489, 492–493
  deviations of 31, 32
  distribution of 26–28
  vs. median 28
  standard error of 38, 39
  X, item "difficulty," and *p* 28–30
mean absolute deviation (MAD) 494–495, 494*t,* 495*t*
mean square regression (MSR) 552–553
mean square total (MST) 528–529, 529*t*
measurement
  approach to 353
  definition of 109–110, 348
  foundations of 19–21
  instruments *107,* 107–108, 154–155
  values 109–110
measurement error
  application process 163
  assessment 157–158
  calibration 162, 163
  concatenation *vs.* instrumentation's indirect results 165–166
  definition 160–161
  derived measurement 166, 167*t*
  design and development resources 158–159
  direct measurement 165
  domain theory 155–157
  ECD 160
  empirical and informational process 161
  Gauss model for replications 164
  indirect measurement 165
  measurement instrument 154–155
  observed score, true score, and error 168–169
  PADI 160
  precision
    European Governmental Standards and Guidelines 171
    imprecision and relative error 169–170
    standardization requirements 169
  reliability
    CTT (*see* classical test theory (CTT))
    Fisher's information and efficiency 173–174
    vs. precision 172–173, *173*
    Shannon's entropy 174–175
  review questions 185–189
  stochastic variables 24, *24,* 25
  test–retest method 168
  thermometers 167
  transduction 162, 163
measurement scales
  in behavioral and social sciences 23
  Ferguson Committee 12–13
  foundations of *16,* 16–21, *20*
  metrology and measurement units 21–22, *23*
  review questions 58
  RTM 13
  scale types 13
  Stevens' four scales 13–16, *14*
  true score theory 20–21
median 485–489
  definition 27
  vs. mean 28
Mendel's qualitative approach 6
Messick, Samuel 100–108
metrology
  applied metrology 327–328
  extensive metadomain 332–334
  history of SI units 341–342
  intensive metadomain 333–334
  international metrology
    Ferguson Committee 320–321
    history of 318–319
    information world 329
    instrument–scale development before unit definition 331–332
    missing in human sciences literature 322
    sociocognitive caution about properties 331
    VIM 329–331
  lack of international metrology citations 339–340

legal metrology 327–328
measurement communities 108–110
measurement units 21–22, *23*
OSN and RSD 340–341
ostensive metadomain
definition 334–335
outline for human science 338–339
performance criterion assessment 335–337
performance demonstration 335
teaching and assessment roles 338
purposes of 316–317
review questions 343–345
scientific metrology
conflict of interest 322
international 323–325
International Metrology Day (20 May 2019) 325
physical artifacts replaced by fundamental constants 325
SI base units 325–327
social and economic challenges 322
send-off 342
testing instrument *107,* 107–108
Mislevy, Robert J. 105–108
mode 485–489
definition 27
physical characteristic of 27–28
model fit assessment 239–240
model–remodel assessment 253, 255–256
modified graded response model (MGRM) 427–428
modified graded response model with common threshold offsets (MGRM-C) 428–429
motor-frequency (MF) ability 251
Mplus commands 313, 350
Mplus output 376
Mplus syntax 368–369, 379, 380
MST. *see* mean square total (MST)
multiconstruct theories 307
bi-factor models 278, *279*
dimensionality and higher-order constructs 280
hierarchical models 278–280, *279*
multiple layers 280–282, *281*
overview 278
multidimensionality 285–286, 286*t,* 287*t*
multinomial distribution probability model 52–56, *53, 55*
multiple regression 111, 115–125, 349
ANOVA table statistics 552–553
calibration 120
composite scores 118–119
correlation matrix 545–546
correl command 537–538
error in prediction 120, 551
Fisher's information and efficiency measurement 125
graphical basis of *538–540,* 538–545, *543–545*
imprecision 123–125
independent variables and sample size 121
labeled data 536–537, 537*t*
overview 535–536
predicted scores 550
proportion of 'explained' variation 122–123
raw scores 119
review questions 556–559
sample size 553–555
slopes calculation
correlation matrix 547–548
inverse matrix 548
predictive equation 550
standardized regression coefficient 548
unstandardized regression coefficient 549
Y-intercept 549
SSE 551, 552
SSR 550–552
SST 552
standardized multiple regression 119–120
statistical significance 120–121

n

natural (mental) ability 6, 7
negative factor loading 480–481
nonlinear transformation function, standardized scores 81–82
normal distribution
computation of 73–75, *74, 75*
cumulative density function 77–80
inverted ogive *72,* 72–73, *74, 75*
latent variable 216–217
local probability independence 217
percentile position of score 80–81

normal distribution (*cont'd*)
rationale 77
standard normal curve *76,* 76–77
normal ogive (probit) IRT models *414,* 414–415
Normed Fit Index (NFI) 240–241

o

ogive curve 78–80, *79*
Omega reliability
factor loadings and residual variances 472–475
Omega hierarchical 471
Omega hierarchical subscale 472
Omega subscale 472
Omega total 471
one-layer theories
factor rotation and simple structure 284–285, 308
ICS 282–283, *283,* 308
unidimensionality *vs.* multidimensionality 285–286, 286*t,* 287*t*
one-parameter logistic (1PL) model 401–402, 412, 416–417
open skills network (OSN) 340–341
ordered polytomous items
category comparison methods 417–418, 418*t*
CIFA 418, 419*t*
common discrimination parameter
with item location and common threshold offsets 422
and step thresholds 421–422
item discrimination parameters
with item location and common threshold offsets 422
and step thresholds 421
model equivalence, nesting, and comparisons 419–421, *420,* 420*t,* 421*t*
model specification 418, 419*t*
original correlation matrix 240
ostensive demonstration
definition 352–353
of skills 355–357
Ostensive Demonstrations 338–339
ostensive metadomain
definition 334–335
outline for human science 338–339
performance criterion assessment 335–337
performance demonstration 335
teaching and assessment roles 338

p

Parkinson's disease 281
partial correlation 128–131, 134
partial credit model (PCM) 425, 430
Pascal, Blaise 50, 51
Pascal's triangle calculator 61–62
path diagram 287, *288*
patterned matrices 288
Pearson product-moment correlation coefficient (PPMCC) 8, 501–502, 522
performance domain 195, 199
Petr Blahuš's perspective 348
physical coordination ability 227
point estimates
population standard deviation 37
probable error 38–40
sample mean 36
unbiased estimates 37, 38
population inference. *see* statistical inference
precision
European Governmental Standards and Guidelines 171
imprecision and relative error 169–170
vs. reliability 172–173, *173*
standardization requirements 169
prescribed structural hypotheses 287
principal component analysis (PCA) 2, 300–306, 312
Principled Assessment Design for Inquiry (PADI) 108
design for 159
documentation and resources 160
principle of local independence 131
principle of local uncorrelatedness 130–131, *131*
principle of partial uncorrelatedness 130
principles of representation/meaningfulness 11
probabilities-weighting method 513–516, 514*t*, *516*
probability density function (PDF). *see* normal distribution
probability model *24,* 24–25
probability theory
independent joint events, multiplication rule *42,* 42–44, *44*

mutually exclusive composite events, addition rule 44–46
principles 40, 42
surveys 40–41
progressive attainment 192
FORE domain 193–194
internal aspects of validity 193
learning progression 194
model-based reasoning 194–195
multiple revisions 193

q

quantification
behavioral sciences 3–4
behavioral statistics 4–5
correlation coefficient 5–9
least squares regression 5–9
standard deviation 65–71
standardized scores (*see* standardized scores)
Quartimax method 284
quasi-quantification 30

r

randomness, measurement error 178–179
Rasch model 401, 405–406, 415, *416,* 417
rating scale model (RSM) 425
raw scores, formulas for variance 510
Reading Comprehension Ability (RCA) 144–146
regression
with correlation and heritability 7–9
historical development 5–6
residual 231
review questions 57
regression-weighted score 118–119
relative error 169–170
relative model fit 240–241
reliability
coefficient Alpha
bootstrapping 376
congeneric data 377–378
estimation of 378–379
Mplus output 376
parameters 375
statistical modeling program 375
coefficient Omega
vs. Alpha 380–382
bootstrapping 380
item deletion 382–383
Mplus syntax 379, 380
coefficient Rho 383–384
confidence intervals 395*t*–396*t*
CTT (*see* classical test theory (CTT))
Fisher's information and efficiency 173–174
formulas 391, 392*t*–393*t*
maximal reliability and optimally weighted indicators 384–385
vs. precision 172–173, *173*
Shannon's entropy 174–175
true score models 360–361
reliability assessment
correlated indicator residuals 481, 481*t*–482*t*
negative factor loading 480–481
reliability estimation, multidimensional constructs
higher-order models, analytic formulas 469–470
reliability index
bifactor CFA 467–469, *468*
for correlated factors model *464,* 464–466, *465*
for higher order factor model 466–467, *467*
stratified alpha 462–464, *464*
reliability information
average relative parameter bias 477–478
construct replicability index 477
factor determinacy 478–480
reliability vs. precision 172–173
representational theories of measurement (RTM)
foundations *16,* 16–21
review questions 58
scale types and 13
reversion line 7
rich skills descriptions (RSD) 340–341
Root Mean Squared Residual (RMSR) 240
row vector 506

s

sampling distribution, variables *24,* 24–25
scalars 506–508
scientific explanation
correlation
vs. causation 125–127
internal and external validity 127–128, 128*t*
partial correlation 128–129

scientific explanation (*cont'd*)
spurious correlation 128–129
"determining" *IQ* in children 129–130
local uncorrelatedness and falsifiability 130–131, *131*
multiple regression 122–125
overview 121–122
sampling stratification 130
statistical control 130
statistical metaphor 129
scientific metrology
conflict of interest 322
international 323–325
International Metrology Day (2019) 325
physical artifacts replaced by fundamental constants 325
SI base units 325–327
social and economic challenges 322
scientific rigor, variables 11
scientific variables, statistical and test theory methods *24,* 24–25
scoring
dispersion of 31–33
quantification (*see* quantification)
review questions 93–96
standard scores 82–91
semantic correspondence rules 226–227
sensorimotor coordination ability 226
Shannon's entropy 174–175
signal detection theory (SDT) 21
S minus 1 (S-1) bifactor model 443, *443,* 444
software tools 350–351
SPSS commands 260–269
squared multiple correlation (SMC) 249–250
standard deviation 497–498
standard deviation (SD). *see also* standardized scores
binary tests 67–68
bivariate regression 521–523
inappropriate scoring 67
overview 65–67
percents *vs. Z*-scores
C-Praxis 70–71
norms and baselines 68–70
V-Fluency 70–71
standardized Alpha 386–387
standardized Omega 387–389
standardized reliability
Coefficient *H* 390–391
standardized Alpha 386–387
standardized Omega 387–389
standardized Rho 389–390
standardized Rho 389–390
standardized scores
change of performance 81–82
compensation 90–91
composite score 88–90
differential weighting 91
floor and ceiling effects, J-shaped and L-shaped distribution *86,* 86–87
IQ, T-score, and stanin scale 83–84
linear transformation 81
nonlinear transformation 81–82
non-normal distributions 87–88
normal distribution (*see* normal distribution)
overview 66
percents *vs.* standard deviation
C-Praxis 70–71
norms and baselines 68–70
V-Fluency 70–71
population selection and extreme groups 88
reference-dependent preferences 85, *85*
skewed asymmetric distribution *84,* 84–85
test profiles 91
standard scores (Z-scores) 499–500, 499*t*
Stata commands 269–274
Stata factormat procedure 311–312
Stata sem procedure 309–311
statistical inference
binomial distribution (*see* binomial distribution)
classical probability theory 40–46
descriptive sample variance 36–37, 37*t*
inferential estimate 36–37
mathematical assumptions 36
multinomial distribution probability model 52–56
overview 35–36
point estimate 36–38
population standard deviation 37–38
probable error 38–40
systematic error 36
statistical methods
behavioral sciences 3–4
example 33–35, 33*t*, *34*, 35*t*
group location 25–28, *26*

intra-group and inter-individual variability 31–33
mean *X,* item "difficulty," and *p* 28–30
median *vs.* mean 28
review questions 57, 59–61
sample and population (*see* statistical inference)
Stevens' four measurement scales 13–16, *14*
Stimulus–Reaction *(S–R)* law 82
stochastic variables 11, *24,* 24–25
strictly congeneric model 363, 371
structural equation modeling (SEM) software 365
sum of squares error (SSE)
bivariate regression analysis 527–528
multiple regression analysis 551, 552
sum of squares regression (SSR)
bivariate regression analysis 526–528
multiple regression analysis 550–552
sum of squares total (SST)
bivariate regression analysis 527–528
multiple regression analysis 552
Suppes, P. 2, 16, 56
syntactic correspondence rules 227–230, *228*
System of Units (SI) 21–22, *23*

t

tapping test 251, 252*t*
tau-equivalent models 361, 370–371
TestComp 350
test–retest analysis 251, 253, 253*t,* 254*t*, *255*
test–retest method 168, 181
Thurstone's Factor Theorem 239
transduction function 229
transduction, measurement error 162, 163
transpose, vector and matrix 506–507
true score models
congeneric model
equally reliable indicators 372–374, 373*t,* 374*t*
Mplus syntax 368–369
factor models 363–365, 363*t, 364,* 364*t*
inter-item and item-total correlations 366–368, 368*t*
item reliability 371–372
latent variable analysis 361–362, *363*
latent variable modeling 365, *365*
multiple items 361
reliability 360–361
strictly congeneric model 371
tau-equivalent and parallel models 370–371
true score theory. *see* classical test theory (CTT)
Tversky, A. 2, 16, 56, 85
two-parameter logistic (2PL) model 402, *402,* 412, 417

u

unbiased variance 498–499
uncertainty 174–175
unidimensional and multidimensional factor analysis
bifactor models *442,* 442–444, *443*
CFA 439, 440, *440*
data example 444–445
EFA 439, *440,* 441, *441*
higher-order factor models *441,* 441–442, *442*
with multiple factors 438–439
unidimensionality 241, 285–286, 286*t,* 287*t,* 294–296
Unified Parkinson's Disease Rating Scale (UPDRS) 281

v

validity
constructs 100, 102
correlations 99–100
definition 99, 101
external
aspects 112
composite validity 115
criterion-related validity 113–114
maximal validity 116–117
multiple correlation 117–118
multiple regression 118–121
overview 111–112
standard error of estimation 114
standardized composite scores 112–113
subtests 116
variability and 115
weighted composite validity 116
internal
causal effects *vs.* independent/dependent variables 136–139, *137, 138*
causal interpretation 139–140
empirical and observable 134–135

validity (*cont'd*)
formalized and unformalized levels 147
latent common factor model 134
latent variables 132–133, 133*t,* 135–136
local uncorrelatedness 132–133, 133*t*
manifest-latent link 142–146, 143*t*
measurement approach 141–142
partial-local uncorrelatedness 135–136
propensity toward behavior *140,* 140–141
theoretical and unobservable 135
material and references 110–111
measurement *107,* 107–110
observations 100
perspectives on 353–355
professional standards 102–103
review questions 148–150
scientific explanation (*see* scientific explanation)
test interpretations *101,* 101–102
test use *101,* 101–102
validity argument 103–107
validity index 99–100
variables
data and 9–11
measurement scales (*see* measurement scales)
with non-normal distributions 87–88
principles of representation/meaningfulness 11
review questions 57–59
scientific *24,* 24–25
scientific rigor 11
stochastic 11, *24,* 24–25
variance 495–497, 503–504
biased and unbiased variance 498–499
of mean 39
population 36, 55–56
of SDM 52–54
standard deviation (*see* standard deviation (SD))
survey 41
Varimax method 284
Verbal Fluency (V-Fluency) 70–71
Vocabulary of Metrology (VIM) 328–331

W

Weak Associative Measurement 228–230
Weber-Fechner law 82
weighted least squares (WLS) fitting 241–242
Western Governors University (WGU) 155, 338, 355–357, 356*t*
Wilson, M. 108–110, 141, 144, 146

Z

Z-scores. *see* standardized scores