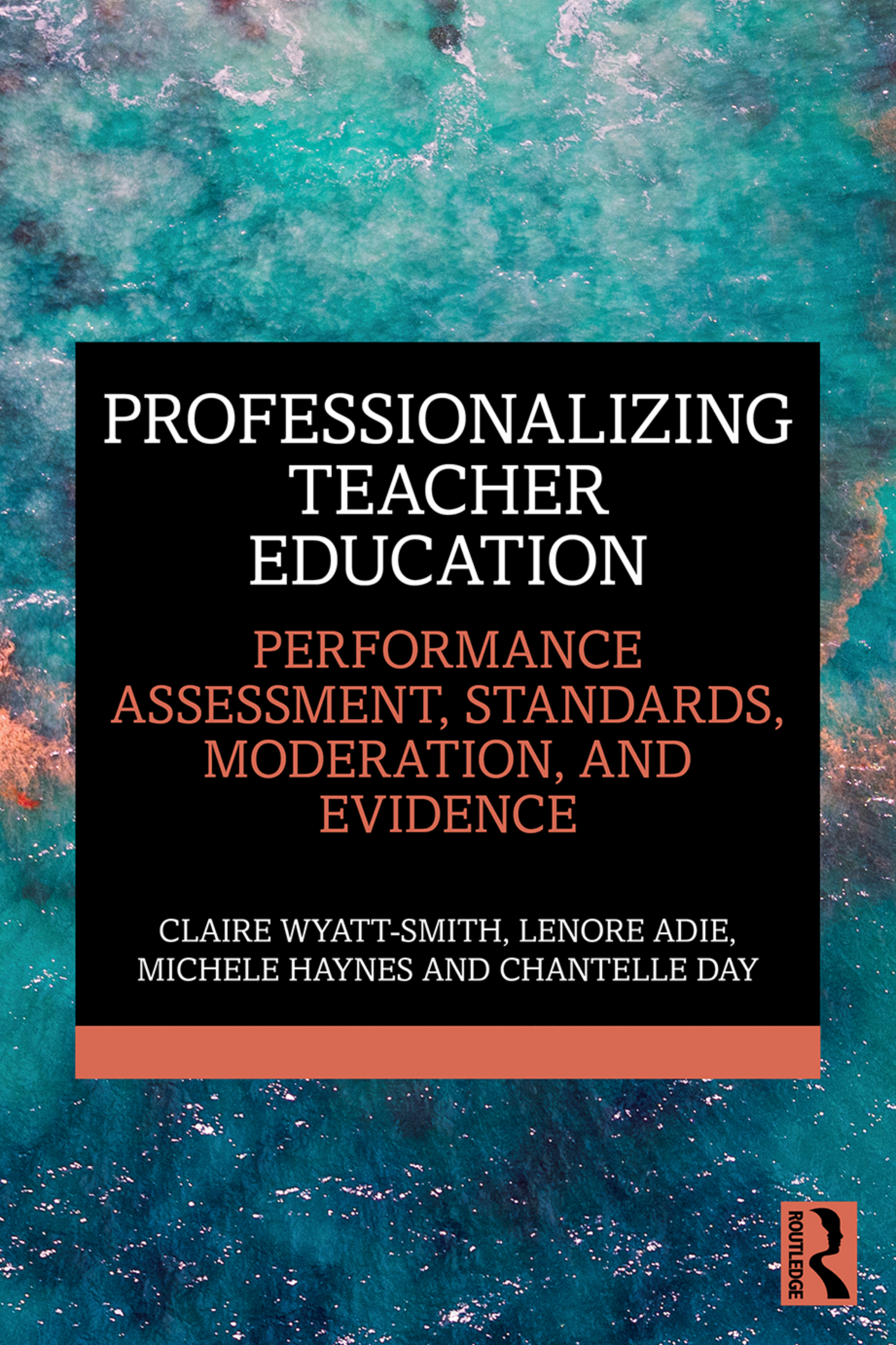


PROFESSIONALIZING TEACHER EDUCATION

PERFORMANCE ASSESSMENT, STANDARDS, MODERATION, AND EVIDENCE

CLAIRE WYATT-SMITH, LENORE ADIE,
MICHELE HAYNES AND CHANTELLE DAY

"*Professionalizing Teacher Education: Performance Assessment, Standards, Moderation, and Evidence* provides original, informative and significant insights into the black box of the design, development, validation, standard setting and implementation of the Graduate Teacher Performance Assessment (GTPA®) in Australia, as well as an account of the innovative and distinctive approach to cross-institutional standards-referenced moderation adopted. As such, this book will be necessary reading for all those interested in authentic assessment practices, in intelligent accountability, in teacher education reform and in improving the status and professionalism of both teacher education and teaching."

Professor Bob Lingard PhD FASSA FAcSS,
Australian Catholic University

"This is an important book, providing an empirical response to the perennial question: 'What's good enough for beginning teachers, and how would you know?' It draws substantially on a major national standard setting, benchmarking, and moderation reform initiative. This is an impressive contribution to the international debate, theoretically grounded, and empirically sophisticated."

Bill Louden AM,
Emeritus Professor of Education,
University of Western Australia

PROFESSIONALIZING TEACHER EDUCATION

This book provides a significant contribution to conversations about teacher quality and graduate readiness for teaching. It presents empirical insights into how a multidisciplinary team of researchers, teacher educators, and policy personnel mobilized for collective change in a standards-driven reform initiative. The insights are research-informed and critically relevant for anyone interested in teacher preparation and credentialing. It gives an account of a bold move to install a collaborative culture of evidence-informed inquiry to professionalize teacher education.

The centerpiece of the book is the use of standards and evidence to show the quality of graduates entering the teaching workforce. The book presents, for the first time, a model of online cross-institutional moderation as benchmarking to generate large-scale evidence of the quality of teacher education. The book also introduces a new conceptualization of a feedback loop using summative data for accountability and formative data to inform curriculum review and program renewal.

This book offers the insider story of the conceptualization, design, and implementation of the Graduate Teacher Performance Assessment (GTPA). It involves going to scale with a large group of Australian universities, government agencies, and schools, and using participatory approaches to advance new thinking about evidence-informed inquiry, cross-institutional moderation, and innovative digital infrastructure.

The discussion of competence assessment, standards, and change processes presented in the book has relevance beyond teacher education to other professions.

Claire Wyatt-Smith (Ph.D.) is Professor of Educational Assessment and Measurement and the Director of the Institute for Learning Sciences and Teacher Education, Australian Catholic University. Her research focuses on standards, moderation, professional judgment, and the implications of digital disruption for teacher professionalism. Current research includes a large-scale Australian study working with a national

Collective of universities on the design and implementation of the Graduate Teacher Performance Assessment (GTPA). She has held leadership roles in universities and schools in Australia and advisory roles internationally. She currently leads a longitudinal quantitative analysis of the quality and impact of initial teacher education. She has undertaken many large-scale studies with Australian Research Council and other research consultancy funds. Her latest books with colleagues are *Digital Disruption in Teaching and Testing: Assessments, Big Data and the Transformation of Schooling* (2021, Routledge) and *Teaching Performance Assessments as a Cultural Disruptor in Initial Teacher Education: Standards, Evidence and Collaboration* (2021, Springer).

Lenore Adie (Ph.D.) is Associate Professor of Teacher Education and Assessment, and Senior Research Fellow at the Institute for Learning Sciences and Teacher Education, Australian Catholic University. Her research focuses on assessment and moderation processes as these contribute to quality assurance and improvement purposes. Her research has generated new knowledge in the field of assessment, focusing on quality in assessment practices and processes, in particular, within systems of standards-referenced assessment. She currently leads an Australian Research Council project investigating the use of scaled annotated exemplars of achievement standards in online moderation to improve teacher assessment capability. She has extensive professional experience working in schools as a teacher and in leadership positions, and in teacher education for over 30 years.

Michele Haynes (Ph.D.) is Professor of Data Analytics for Education Research at the Institute for Learning Sciences and Teacher Education, Australian Catholic University. She is an accredited statistician in Australia with extensive experience as a statistical methodologist and innovator using longitudinal data for education research. Michele has expertise in the estimation of complex models for social applications using data from multiple sources, including panel surveys and administrative data.

Chantelle Day (Ph.D.) is Research Partnerships Manager at the Institute for Learning Sciences and Teacher Education, Australian Catholic University. A significant priority within her portfolio includes management of the Institute's largest, longitudinal research project titled the Graduate Teacher Performance Assessment: Standards and Moderation Project. Chantelle has experience in managing and supporting research projects involving various stakeholders and has worked at the Institute since the completion of her doctoral studies in 2017. Chantelle's research expertise extends to the fields of equity and inclusion in higher education, teacher education, and assessment.

PROFESSIONALIZING TEACHER EDUCATION

Performance Assessment, Standards, Moderation, and Evidence

Claire Wyatt-Smith, Lenore Adie,
Michele Haynes, and Chantelle Day

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Cover image: © Abstract Aerial Art/Getty Images

First published 2022
by Routledge
4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record has been requested for this book

ISBN: 978-0-367-33213-6 (hbk)
ISBN: 978-0-367-33212-9 (pbk)
ISBN: 978-0-429-31850-4 (ebk)

DOI: 10.4324/9780429318504

Typeset in Bembo
by SPi Technologies India Pvt Ltd (Straive)

CONTENTS

FIGURES

TABLES

FOREWORD

Conceptualizations and representations of the *classroom readiness* of newly qualified teachers are much discussed by policy-makers, teacher educators, and school leaders. Indeed, given the importance of teacher quality to the success of school systems, these representations are of wider public interest. The development of the Graduate Teacher Performance Assessment (GTPA) in Australia may have been initiated by the heated public debate about teacher quality, but this book is an account of the project led by researchers at Australian Catholic University's Institute for Learning Sciences and Teacher Education that was informed by considerations of collective and professional agency rather than by simplistic rhetoric of "testing" and "checking" teachers on exit from their preparation programs.

The identification of the five core practices – *Planning*, *Teaching*, *Assessing*, *Reflecting*, and *Appraising* – that underpin the GTPA will be of interest to teacher educators and researchers in that field. They present the heart of teaching as attention to learning in two ways. First, they position the teacher as a learner, coming to an understanding of the impact of their teaching on what and how students learn and how they progress. Second, they make explicit connections between the informed actions of the teacher and the student learning that may or may not follow. These core practices call attention to the differences between intended and enacted classroom practice; the lesson plan that provides the structure, delivery focus, and scaffolding for many students and newly qualified teachers across the globe is replaced in the construction of this GTPA by a more complex process, in which planning continues to play a key role, but is secondary to impact and outcomes.

The Australian context adds further depth to the account; the initial call by the government for the development of a teaching performance assessment was for consortia of institutions to come together and provide opportunities for graduating teachers to demonstrate their professional competence. Such collaborations on academic standards and shared understandings of quality are rare in higher education more broadly, and the processes by which these were moderated and finalized are described in some detail. In some ways, the rigorous attention to evidence demonstrated in the development of the GTPA models the rigor expected of student teachers in engaging with data in all its forms as part of their assessment.

The development and deployment of this GTPA is of interest beyond teacher education; those with an interest in the processes by which all professionals are educated, credentialed, and licensed will find a rich resource in this publication.

Professor Anne Looney
Executive Dean | DCU Institute of Education | Dublin City University

ACKNOWLEDGMENTS

In this book, we share our journey into the world of teaching performance assessments (TPAs). The journey has taken us beyond the design of a competence assessment. We have been guided by the informing literature, research, evidence, and voices in the field. We have been inspired through the dynamics of the collaboration. In our exploration, we have navigated new methodologies for collecting and analyzing data, standard setting in teacher education, interdisciplinary teamwork, and going to scale.

We have also been inspired by the invited contributions of our esteemed colleagues, Honorary Professor Joy Cumming (Australian Catholic University) and Professor Emerita Diana Pullin (Boston College). Their writing for this book addresses some of the most significant educational issues of our time, including fairness in assessment, educational testing and the law, and the use of standards and competence tests. The book is richer for their contributions and insights into teacher education reform and legal issues of significance to the introduction of teaching performance assessments in Australia.

We thank the Australian Institute for Teaching and School Leadership (AITSL) for the initial funds through the Teaching Performance Grant Program ($AUD350K). This was to support the one-year Trial and included the conceptualization, development, and realization of the Graduate Teacher Performance Assessment (GTPA). We give special thanks to the Australian Catholic University and the Office of the Deputy Vice-Chancellor (Research) for sustained support of the GTPA Project to build a national evidence base in Australian initial teacher education.

The project has been complex and intense, involving many elements and stakeholders. A key learning has been that in order to consider teacher education research as nation-building, new ways of thinking and acting are necessary. This involves bringing together universities, regulatory authorities, schools, employing authorities, and unions into a model for collective responsibility and action to generate evidence of quality and effectiveness.

Throughout the journey, the shared priority of preparing high-quality teacher education graduates for Australian classrooms has been clear and sustained. While success can be measured in many ways, one notable achievement is how the GTPA Collective of universities nationwide has mobilized the largest, cross-state body of teacher educators, taking agency to develop a national understanding of the characteristics of quality graduates.

Special appreciation goes to the Collective (see https://www.graduatetpa.com/discover/ for a list of participating universities). We recognize the impact of their contributions to knowledge generation in teacher education. For many, engagement has been sustained since the 2016 pilot and 2017 trial. We have vivid recollections of rich professional discussions about the instrument and use of technologies that have shaped how the Collective works today. The willingness of teacher educators to ask critical questions about standards, evidence, and preparedness for teaching has been inspirational. Undoubtedly, they have applied their expertise and agency to improving initial teacher education.

Thanks go to the GTPA reference group of internationally recognized teacher education scholars and the GTPA Advisory Board members who represent 11 peak national bodies. From the beginning, they have provided research-informed and strategic policy advice to inform the GTPA trial and implementation over several years.

We are also indebted to the thousands of preservice teachers who completed the GTPA as part of their program requirements. Without their submissions and willingness to support research into graduate readiness, the GTPA research project would not have been realized.

Finally, thanks go to the GTPA multidisciplinary research team whose work contributed to this book: Dr Melanie Spallek, Dr Andrew Smith, Mr Alex Chen, Mr Alex Mason, Ms Famena Staley, and Dr Elizabeth Heck. We recognize Melanie's key role in data analytics and presentation. Andrew has made significant contributions to the systems thinking in the GTPA project. We are grateful to Alex Chen for his knowledge, skills, and patience in managing the technologies used in this complex project. We also recognize Alex Mason for his demonstrated expertise in visual design and his artistry in video production to support the GTPA Collective. We acknowledge the contribution of Famena for her sustained engagement in the preparation of the manuscript. Finally, in her role as editorial manager, Elizabeth has demonstrated considerable expertise in manuscript development and preparation. A special mention is also made of the contribution of Peta Colbert in the Pilot and Trial of the assessment.

CONTRIBUTORS

Diana C. Pullin (J.D., Ph.D.) is Professor Emerita at the Lynch School of Education and the School of Law at Boston College. She is a leading U.S. and international expert in law and assessment, testing and accountability systems for teacher education, performance, and certification. She has published extensively on education law and public policy; testing and the law; educational leadership and teaching, and the impact of social science on legal decisions in education. Other areas of expertise include opportunity to learn; educator quality; accountability in higher education; and impact of social science research on legal and public policy decisions in education. Professional standards of practice have also been a focus of her work; she is one of the co-authors of the 1999 *Standards on Educational and Psychological Testing* and she served as well for a number of years as a member of the Joint Committee on Standards for Educational Evaluation.

Joy Cumming (J.D., Ph.D.) is Honorary Professor in the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University, where she was Research Director of the Assessment, Evaluation and Student Learning research area from 2013, retiring in 2019. Previously she has been a secondary school teacher, researcher, and academic at Griffith University holding major administrative positions. Her nearly 50 year career in education research has focused on policy and practice in literacy and numeracy for all age levels, and assessment, incorporating international and national funded projects in these areas, and resulting in a large corpus of written reports, books, and journal papers. Her research work in assessment is strongly focused on equity, especially for students with disability, and design of assessment processes that fairly represent what students know and can do. Qualified as a lawyer in 2009, her assessment equity research takes both a legal and social justice perspective.

ABBREVIATIONS

AACTE	American Association of Colleges for Teacher Education
ACARA	Australian Curriculum, Assessment and Reporting Authority
ACER	Australian Council for Educational Research
ACU	Australian Catholic University
AERA	American Educational Research Association
AHPRA	Australian Health Practitioner Regulation Agency
AITSL	Australian Institute for Teaching and School Leadership
ANSAT	Australian Nursing Standards Assessment Tool
APA	American Psychological Association
APCAT	Australasian Paramedic Competency Assessment Tool
APST	Australian Professional Standards for Teachers
ATAR	Australian Tertiary Admissions Rank
BTL	Bradley–Terry–Luce Model
CIM	Cross-Institutional Moderation
CIM-ONLINE™	Cross-Institutional Moderation – Online
CTH	Commonwealth
DPJM	Dominant Profile Judgment Method
DSE	Disability Standards for Education
EDTPA	Educative Teacher Performance Assessment
EJTE	*European Journal of Teacher Education*
EQuITE	Evidence for Quality in Initial Teacher Education
FERPA	Family Educational Rights and Privacy Act
GPA	Grade Point Average
GTPA	Graduate Teacher Performance Assessment
ILSTE	Institute for Learning Sciences and Teacher Education
IRM	Item Response Models
ITE	Initial Teacher Education

JCU	James Cook University
LANTITE	Literacy and Numeracy Test for Initial Teacher Education
LSIA	Learning Sciences Institute Australia (now ILSTE)
MFRM	Multi-Facet Rasch Model
NAPLAN	National Assessment Program – Literacy and Numeracy
NCME	National Council on Measurement in Education
NESA	New South Wales Education Standards Authority
NMBA	Nursing and Midwifery Board of Australia
NSW	New South Wales
NT	Northern Territory
OFSTED	Office for Standards in Education
OP	Overall Position
OTD	Overseas Trained doctor
PACT	Performance Assessment for California Teachers
PAT M	Progressive Achievement Test – Mathematics
PAT R	Progressive Achievement Test – Reading
PISA	Programme for International Student Assessment
PLDS	Performance Level Descriptor(s)
QCAT	Queensland Civil and Administrative Tribunal
QCT	Queensland College of Teachers
QLD	Queensland
RMIT	Royal Melbourne Institute of Technology
SA	South Australia
SCALE	Stanford Center for Assessment, Learning and Equity
TEMAG	Teacher Education Ministerial Advisory Group
TPA	Teaching Performance Assessment
UK	United Kingdom
UQ	University of Queensland
U.S.	United States
UTAS	University of Tasmania
VIC	Victoria
WA	Western Australia

PART 1

Conceptualization, design, and implementation

1

STANDARDS, LARGE-SCALE EVIDENCE, PROFESSIONAL JUDGMENT, AND THE AFFORDANCES OF DIGITAL TECHNOLOGIES IN TEACHER EDUCATION

Teacher education and issues of evidence and quality

This book tells the story of a reform initiative in initial teacher education (ITE) aiming to prepare 'classroom ready teachers': What the introduction of a teaching performance assessment (TPA) actually looks like and how it stimulates culture change. The book takes up three questions. *How did change occur? How much progress has occurred in reforming ITE using TPAs in the last five years? What more needs to be done to install an agreed standard for graduate teachers on entering the profession?*

Currently, societies are being formed and reformed within a crucible of change triggered by geo-political and socio-economic conditions, new and emerging technologies, climate-related changes, and global pandemics. The recurring message across the global community is that we are living in unprecedented times. This has intensified calls for thinking differently about the purposes and practices of education and schooling systems: How can we educate for an informed citizenry, high levels of literacy and numeracy, and radically different workplaces where the knowledge, skills, dispositions, and communicative competences that had currency in preceding industrial eras are no longer relevant or sufficient? Schwab (2017) identified four stages of industrial change ranging from steam, science, digital technologies, and now into the era of increasing computer power and data including the use of cloud and mobile technologies, and artificial intelligence.

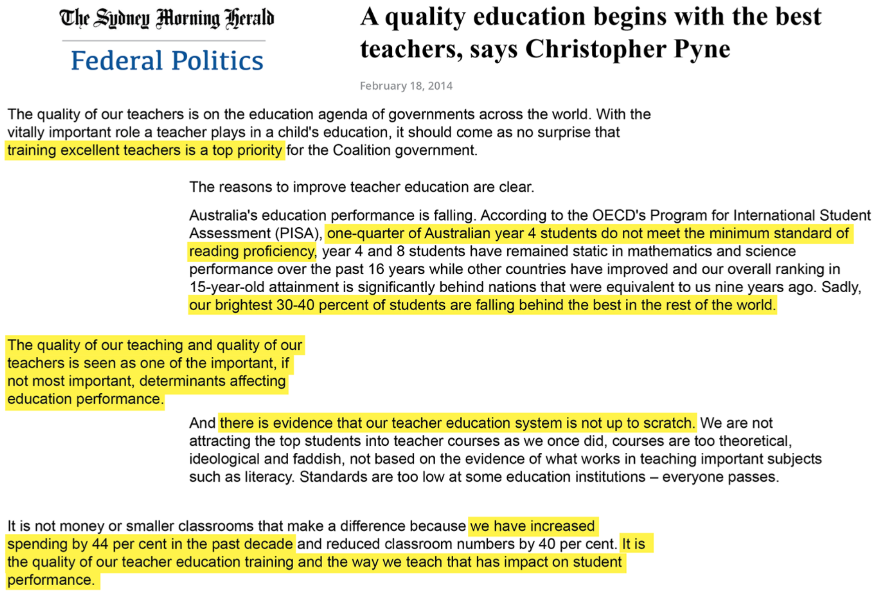
In this broad context, teacher education and assessment have been recognized as fields for much-needed reform. This reflects how assessment evidence, including test data, has been highly valued for its function in reporting student achievement. However, it has been underutilized to investigate how student learning occurs in real time and how progression occurs over time. This is the case even though the assessment for learning movement has been recognized for some time as influential in several countries including the UK, Canada, New Zealand, and Australia

DOI: 10.4324/9780429318504-2

(Broadfoot & Black, 2004). Here, we reflect how the traditionally recognized purposes of assessment – formative (improvement), summative (measurement) – have tended to operate on separate fronts. We also observe how school assessment practices have been resistant to change since the introduction of mass schooling, though COVID-19 has challenged some of the long-standing assumptions about 'good' assessment practices (Broadfoot, 2007; Lingard et al., 2021).

Calls for reforming teacher education have been growing in many countries with the print media capturing public dissatisfaction with the quality of graduates entering the teaching profession. This has occurred at a time of intensifying attention given to the results of international tests that are used to serve government interest in gauging the quality of schooling systems. UNICEF's joint statement on World Teacher's Day emphasized the criticality for ensuring quality teachers and quality teacher education as a basic right of all children to secure a better future (*Targeted News Service*, 2018). Globally, concerns with teacher quality and efforts to improve teacher quality are evident. For example, the introduction of teacher professional standards in South Africa aimed to improve teacher quality and professionalism, with ITE reported as one factor contributing to the poor quality of teachers (*Businge*, 2019; Robinson, 2019). This resulted in major changes for teacher education in the country (Wanzala, 2019). In New Zealand, there has been media commentary that teacher education was failing to produce quality teachers (e.g., Collins, 2017) with some identifying systemic weaknesses in teacher preparation (e.g., Jones, 2017; Moir, 2017; *The Southland Times*, 2017). Similarly, in Canada, teacher education has been reported as needing improvement (Waugh, 2020). In the UK, concern with the quality of new teachers (Denholm, 2017; Wightwick, 2017) has also been raised in the media. For example, attention in Scotland has focused on poor literacy and numeracy skills of trainee teachers (Grant, 2017), the apparent lack of training that student teachers receive in teaching literacy and numeracy (e.g., Johnson, 2017), and a wider call for radically rethinking teacher education (Drew, 2018). Added to these concerns is the pressing issue of teacher shortages as well as decreasing numbers of candidates choosing to study teacher education evident in the literature and the media (e.g., Cochran-Smith, 2020; García & Weiss, 2019; Henebery, 2020; See & Gorard, 2020; UNESCO Institute for Statistics, 2016; Wiggan et al., 2020). This issue is one of the reported triggers for the 2021 review of teacher education in Australia, discussed later in this chapter.

Australian media reports (e.g., Clark, 2017; O'Flaherty, 2020) have sustained the theme of reforming teacher education. For example, O'Flaherty's 2020 report titled "Low OP[1] hurdle dumbing down future teachers" referred to the low academic results required for entry to a teacher education program, and so highlighted a lack of 'quality' candidates in teacher education. This refrain is similarly evident in a previous commentary in the Australian media by the then Federal education minister, Christopher Pyne (2014; Figure 1.1). In this figure, the segments show the reported direct association between falling education performance in the country and "the quality of our teaching and quality of our teachers", described as "one of the important, if not most important, determinants affecting education



The Sydney Morning Herald

Federal Politics

A quality education begins with the best teachers, says Christopher Pyne

February 18, 2014

The quality of our teachers is on the education agenda of governments across the world. With the vitally important role a teacher plays in a child's education, it should come as no surprise that training excellent teachers is a top priority for the Coalition government.

The reasons to improve teacher education are clear.

Australia's education performance is falling. According to the OECD's Program for International Student Assessment (PISA), one-quarter of Australian year 4 students do not meet the minimum standard of reading proficiency, year 4 and 8 students have remained static in mathematics and science performance over the past 16 years while other countries have improved and our overall ranking in 15-year-old attainment is significantly behind nations that were equivalent to us nine years ago. Sadly, our brightest 30-40 percent of students are falling behind the best in the rest of the world.

The quality of our teaching and quality of our teachers is seen as one of the important, if not most important, determinants affecting education performance.

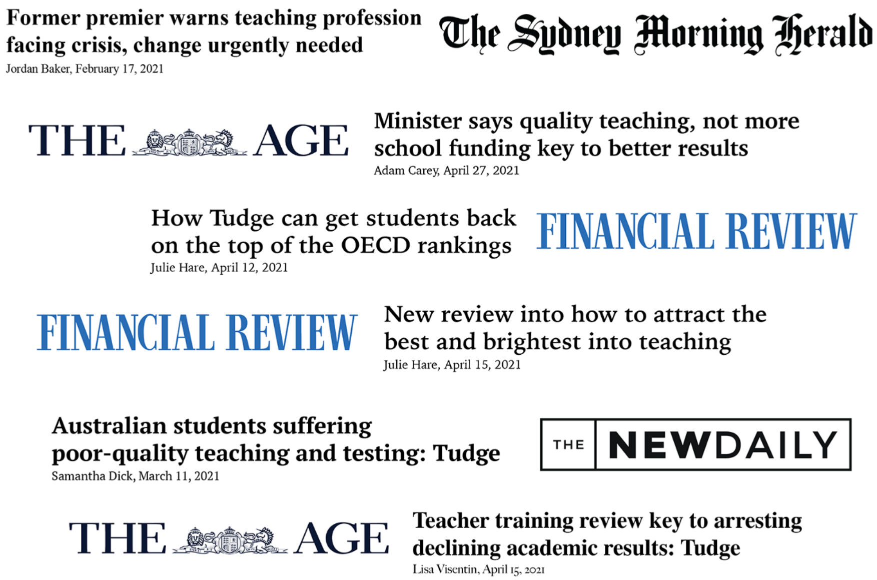
And there is evidence that our teacher education system is not up to scratch. We are not attracting the top students into teacher courses as we once did, courses are too theoretical, ideological and faddish, not based on the evidence of what works in teaching important subjects such as literacy. Standards are too low at some education institutions – everyone passes.

It is not money or smaller classrooms that make a difference because we have increased spending by 44 per cent in the past decade and reduced classroom numbers by 40 per cent. It is the quality of our teacher education training and the way we teach that has impact on student performance.

FIGURE 1.1 Selected extracts from *The Sydney Morning Herald*, Federal Politics (February 18, 2014): Christopher Pyne's comments on quality education and best teachers

performance" (para. 4 in source article). The concerning claim that "one-quarter of Australian year 4 students do not meet the minimum standard of reading proficiency" in PISA (para. 3 in source article), is further strengthened by the statement that Australia's "brightest 30–40 percent of students are falling behind the best in the rest of the world". These performance observations lead to the summary conclusion that teacher education in Australia is "not up to scratch" (para. 7 in source article) and further, that teacher education programs are not attracting the top students as they once did. The connection between performance outcomes and quality is made compelling through the claim of increased spending in education and reduced classroom numbers.

Fast track to 2021, media coverage of teacher education continues to circle around issues of claimed "poor-quality teaching and testing" (see Figure 1.2, *The New Daily* headline) and the push to "get students back on the top of the OECD rankings" (see Figure 1.2, *Financial Review*). Against the background that we have sketched to this point, the Australian public was well prepared for the announcement of a further review into teacher preparation, called by the current Federal education minister, Alan Tudge (2021). This review will extend to "how to attract the best and brightest into teaching" (see Figure 1.2, *Financial Review*) and address the reported overemphasis on theory at the expense of practice and attention to evidence-based teaching methods. Goss et al. (2019) had similarly identified that a way to improve "the quality of the future teaching workforce is to encourage many



The Sydney Morning Herald

Former premier warns teaching profession facing crisis, change urgently needed

Jordan Baker, February 17, 2021

THE AGE

Minister says quality teaching, not more school funding key to better results

Adam Carey, April 27, 2021

How Tudge can get students back on the top of the OECD rankings

Julie Hare, April 12, 2021

FINANCIAL REVIEW

FINANCIAL REVIEW

New review into how to attract the best and brightest into teaching

Julie Hare, April 15, 2021

Australian students suffering poor-quality teaching and testing: Tudge

Samantha Dick, March 11, 2021

THE NEWDAILY

THE AGE

Teacher training review key to arresting declining academic results: Tudge

Lisa Visentin, April 15, 2021

FIGURE 1.2 Collage of 2021 headlines from *The Sydney Morning Herald*, *The Age*, *Financial Review*, and *The New Daily* related to teacher quality in Australian media

more high achievers to apply" (p. 8) with workforce planning identified as necessary to address workforce shortages (Patty, 2021).

Tudge characterized the review as leading to "The next evolution of reforms … to build from the TEMAG [Teacher Education Ministerial Advisory Group] reforms" (para. 71). In his recent speech, he indicated that "This review will investigate where there is still further work to do to ensure that all ITE courses are high-quality and adequately prepare our teachers to be effective from day one" (para. 71). Here, the continuing focus on preparing classroom ready teachers, that came to the fore in the TEMAG reforms, is expected to continue. However, the new refrain is the reference to reforms needed for "arresting declining academic results" (see Figure 1.2, *The Age*) with Tudge identifying three areas of reform: "quality teaching, particularly initial teacher education, curriculum and assessment" (Tudge, 2021, para. 51).

The value of 'quality' teachers and 'quality' teaching is widely advocated as essential for student learning and achievement. Yet the term 'quality' itself is opaque and open to a variety of interpretations. However, if we look across professions, quality and quality performance have some demonstrated characteristics or recognizable features. In medicine, for example, patients discuss looking for the 'best' doctor, referring to aspects such as the effectiveness of treatment and bedside manner; in golf, quality performance is indicated by technical features including golf swing, choice of clubs, difficulty of the course, and the final score card. Irrespective of the field, demonstration of quality performance appears to be bound up with expertise – expert ways of doing, being, thinking, and interacting. Expert performance appears to entail not only recognizable knowledge and skills, but also values and dispositions. The development of expertise or 'quality' performance is recognized as a progression from novice to expert in which performance becomes intuitive rather than the rigid following of rules or steps (Adie et al., 2020). Throughout this book, we return to 'quality' as a central motif. We aim to explore with readers what evidence of 'quality' looks like in the teaching profession, through theory, research, and practice lenses.

In addition to the main questions introduced earlier, the book considers: *What do we know about the quality of beginning teachers' preparedness for practice at the point of entry to the profession? What can we say about the expected professionalism of these teachers?* Currently, data on beginning teachers' preparedness tend to be limited to small-scale studies, typically reliant on qualitative analysis of perceptions and self-reports. To date, scant attention has been given to sustained longitudinal research of actual practice and performance evidence. In Australia, Green et al. (2018) noted that "investigations into the elements of teacher preparation programs that effectively prepare graduate teachers for the realities of the teaching profession are lacking" (p. 104). The authors characterized research into program effectiveness as "invaluable for practitioners and policy makers, particularly in light of the 'reality shock' often encountered by beginning teachers due to a disparity between their tertiary experiences and the classroom realities, contributing to a high early career attrition rate" (p. 105). In New Zealand, a discussion paper into the future options for ITE (Education Council of New Zealand, 2016) identified that

> the evidence base for assessing current ITE provision is relatively thin. We have no data on the actual capabilities of graduating teachers. The nearest we have to any information about this are surveys of graduate satisfaction with their programmes of study…
>
> *(p. 7)*

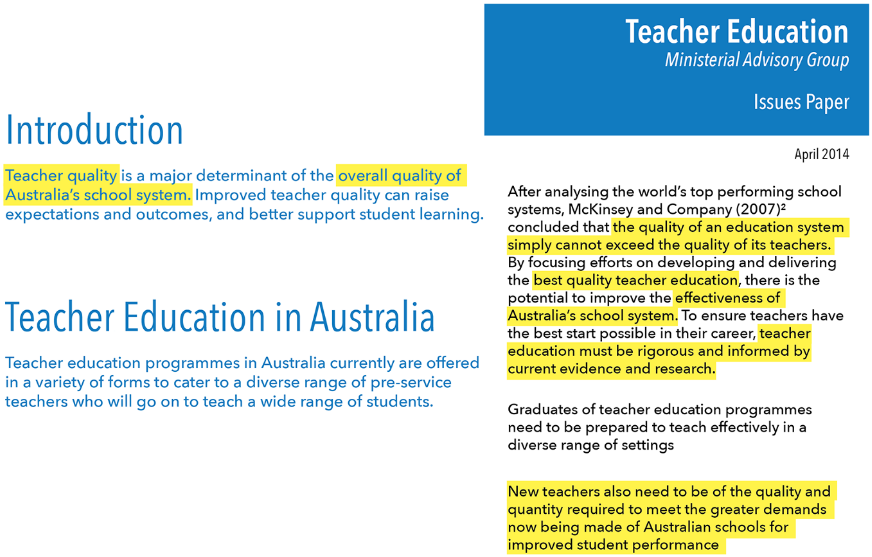
It is widely recognized that teacher preparation routinely involves teacher educators and other stakeholders, including mentors and preservice teachers, using a range of evidence types in the academic program and school-based practical program. However, as mentioned, teacher education as a field lacks a strong evidentiary base to show the quality of teacher preparation and graduate competence on course completion and subsequent entry into the profession (Ell et al., 2019; Rauschenberger et al., 2017; Yeigh & Lynch, 2017). This has left the field open to a succession of reviews of ITE completed during the last decade, and persistent attacks on the status of the profession, including in the media.

A turntable of reviews into teacher education

Across most reviews of ITE is the focus on improving educational opportunities and learning outcomes for school students. Also evident are clear concerns about the relationship between the academic program in universities and the school-based program, and the preparedness of graduates entering the teaching workforce. This is evident in commentary on the most recent review in England: "Supporting our teachers with the highest-quality training and development is the best way we can improve pupil outcomes, and we want all teachers to have a world-class start to their career" (Adams, 2021, para. 13). The sentiments echo across the turntable of reviews in many countries. These include Scotland (Donaldson, 2010), the Republic of Ireland (Sahlberg, 2012), the United States (Cochran-Smith et al., 2013; Rickenbrode et al., 2018), Northern Ireland (Sahlberg et al., 2014), England (Bauckham et al., 2021; Carter, 2015), Wales (Furlong, 2015), New Zealand (Education Council of New Zealand, 2016), Norway (Advisory Panel for Teacher Education, 2020), and Australia (Craven et al., 2014a), with a subsequent review of Australian ITE to be delivered in 2021. This adds to the more than 100 reviews of ITE in Australia between 1979 and 2008 (Louden, 2008).

Broadly speaking, reviews address three main issues: Who is responsible for teacher education? What makes a quality teacher? What mix of elements is necessary to lift the quality of teacher preparation? Referring to Wales, Furlong (2015) concluded that "overall [teacher education] is not of sufficient high quality to serve the needs of Wales either now or in the future" (p. 5) and identified a "lack of leadership of the sector" (p. 11). In New Zealand, the Education Council[2] (2016) referred to "the network of ITE provision [as] uncoordinated" (p. 2). In England, the most recent review seeks to shift "responsibility for ITE away from universities and towards schools" (Clarke & Parker, 2021, para. 4). This is the case, even though the Department of Education in England has provided evidence of the soundness of universities being responsible for ITE (100% undergraduate courses; 70% postgraduate courses) and Ofsted, the UK government quality assurance agency, has rated all ITE institutions as good or outstanding (Clarke & Parker, 2021).

Also live internationally, are the issues of how preparation programs use evidence (a) in the design of programs and in their implementation, and (b) in demonstrating impact on graduate competence in teaching and assessment. Here, we return to the experience in England where there is a clear push back from university providers, including the University of Cambridge and the University of Oxford, regarding the research evidence drawn on to inform significant proposed changes (Clarke & Parker, 2021). These include a move for teacher 'training' in schools as distinct from 'education' in universities. The distinction between training and education for teacher preparation is not a trivial one. There is also the push for standardizing curriculum, the latter impacting on university independence. The intensity of the push-back is evident in the reported response from the University of Cambridge that "it will cease teacher training courses if the government persists with damaging proposals to change how primary and secondary school teachers are trained in England" (Adams, 2021, para. 1), and from the Universities' Council for the Education of Teachers who characterize the latest review as having the potential to destabilize the sector. This shows the risk of what Clarke and Parker (2021) refer to, as "diluting the intellectual standing of the profession" (para. 6).



Teacher Education
Ministerial Advisory Group

Issues Paper

April 2014

Introduction

Teacher quality is a major determinant of the overall quality of Australia's school system. Improved teacher quality can raise expectations and outcomes, and better support student learning.

Teacher Education in Australia

Teacher education programmes in Australia currently are offered in a variety of forms to cater to a diverse range of pre-service teachers who will go on to teach a wide range of students.

After analysing the world's top performing school systems, McKinsey and Company (2007)[2] concluded that the quality of an education system simply cannot exceed the quality of its teachers. By focusing efforts on developing and delivering the best quality teacher education, there is the potential to improve the effectiveness of Australia's school system. To ensure teachers have the best start possible in their career, teacher education must be rigorous and informed by current evidence and research.

Graduates of teacher education programmes need to be prepared to teach effectively in a diverse range of settings

New teachers also need to be of the quality and quantity required to meet the greater demands now being made of Australian schools for improved student performance

FIGURE 1.3 TEMAG Issues Paper (April 2014, pp. 5, 8): Teacher quality (highlighting added)

The need for quality teacher education was also evident in the Australian Teacher Education Ministerial Advisory Group (TEMAG) Issues Paper (Figure 1.3)[3]. Here again, the spotlight is on how "new teachers also need to be of the quality and quantity required to meet the greater demands now being made of Australian schools for improved student performance" (Craven et al., 2014b, p. 5). This was supported by McKinsey and Company's (2007) analysis of the world's top-performing school systems that concluded that "The quality of an education system cannot exceed the quality of its teachers" (p. 40). This is a significant challenge, noting the shortage of teachers experienced in many countries and the widely reported phenomenon of reliance on out-of-field teachers (teaching in areas other than those in which they were prepared).

A key finding of the TEMAG review was the "need to lift public confidence in initial teacher education – Australians are not confident that all entrants to initial teacher education are the best fit for teaching" (Craven et al., 2014a, p. xi). The TEMAG review further asserted that "high-quality teaching is fundamental to student learning, and the biggest in-school factor determining student outcomes" (p. 1). This is a position also highlighted by several authors including Hattie (2003), and Caena (2014) who claimed that "teachers are widely recognized in the research as the most powerful determinants of pupil achievement" (p. 2). There can be no doubt about the need for priority policy attention to be given to quality teaching and preparing quality teachers for the workforce, as concluded in reports from the Grattan Institute (e.g., Goss, 2017; Goss et al., 2017; Goss et al., 2019) and the European Commission (2013).

A troubling finding in international reviews is the lack of preparation of pre-service teachers in assessment and evaluation practices and the use of standards. Furlong (2015), in his review of teacher education in Wales, noted that "there should be a much greater emphasis on teacher led assessment than at present" (p. 7). Referring to teacher education in Scotland, Donaldson (2010) commented that "A further frequent concern was lack of confidence and skills in assessment, including understanding standards and expectations and being able to engage with Scottish Qualifications Authority assessment processes" (p. 35). In this comment, Donaldson (2010) is referring to achievement standards and teachers' skills in understanding expected quality when making judgments of student work. A recent study that reviewed ITE programs in Queensland, Australia, iterated these concerns, concluding that the preparation of teachers to be assessment capable and able to use evidence to inform teaching and improve learning was largely underdeveloped (Wyatt-Smith et al., 2017).

Concurrent with the succession of reviews of teacher education has been the growing interest in competence requirements that all teachers need. In some countries, this has been evident in concerted development work on professional standards for teaching at national and regional levels. In others, there have been attempts to formulate core competence requirements. The European Commission (2013), for example, identified common features across core competences that all teachers need including the following:

- Sound knowledge frameworks (e.g., about school curricula, education theories, assessment), supported by effective knowledge management strategies.
- A deep knowledge of how to teach specific subjects, connected with digital competences and students' learning.
- Classroom teaching/management skills and strategies.
- Interpersonal, reflective, and research skills, for cooperative work in schools as professional communities of practice.
- Critical attitudes toward their own professional actions, based on different sources – students' outcomes, theory, and professional dialogue – to engage in innovation.
- Positive attitudes to continuous professional development, collaboration, diversity, and inclusion.
- The capability to adapt plans and practices for contexts and students' needs.

This formulation highlights the mix of "the intellectual, cognitive and emotional demands of teacher preparation [which] can often appear formidable to student teachers" (Caena, 2014, p. 3).

One of the continuing challenges that is inherent in the reviews and reports relates to distinguishing *teaching* competences and *teacher* competences. The European Commission (2013) distinguishes teaching competences as "the role of the teacher in the classroom, directly linked with the 'craft' of teaching – with professional knowledge and skills mobilised for action" (p. 10). Teacher competences "imply a wider, systemic view of teacher professionalism, on multiple levels – the individual, the school, the local community, professional networks" (p. 10). The Commission further identified that while "dispositions are fundamental for both competence sets, they play a decisive role for teacher competences, embracing attitudes to constant professional development, innovation and collaboration" (p. 10). Noting the distinction between these terms and how they "overlap and interweave" (p. 10) regarding required competences, our interest lies in teaching competence and professional preparedness for practice.

Not surprisingly, the mix of reviews and reports into teaching competence has the potential to affect public confidence in, and the status of, the teaching profession. In turn, they can have significant implications for teacher recruitment and retention. The policy responses across countries reflect the recognition that teacher/teaching quality is consequential for student learning outcomes and also for national rankings on international tests, mentioned earlier. For example, the phenomenon of 'PISA shock' was evident in countries such as Germany, Norway, and, to a lesser extent, Sweden (Haugsbakk, 2013; Volante et al., 2020; Waldow, 2009) when the outcomes showed that international positioning was not as high as anticipated. As a result, there was great public alarm regarding the quality of the nation's education system with consequential vast educational reforms.

In many countries, reviews into teacher education and the quality of teaching, together with published outcomes of international testing, have fueled changes in education policy and a suite of initiatives intended to achieve excellence in schooling. Examples include the *Review to Achieve Educational Excellence in Australian Schools* (Gonski et al., 2018) that examined the challenges of achieving both equity and excellence in the education of the nation's young people. This review resulted in changes to the models of school funding in Australia. Echoes of equity and excellence reforms can be heard in several other countries, though relative emphases and implementation procedures and processes vary significantly. To improve schooling in the United States, *Race to the Top* (U.S. Department of Education, 2016) identified four key areas of reform:

- Development of rigorous standards and better assessments.
- Adoption of better data systems to provide schools, teachers, and parents with information about student progress.
- Support for teachers and school leaders to become more effective.
- Increased emphasis and resources for the rigorous interventions needed to turn around the lowest-performing schools.

Two additional examples further illustrate how countries are seeking to address issues of teaching quality, though through different approaches. First, drawing on another example from the United States, Crowe (2011) identified that to improve the quality of schooling, it is essential to "measure student gains and associate student achievement with specific teachers; and link teachers to their teacher preparation programs" (p. 25); the latter intended to gauge the effectiveness of teacher preparation. Second, in the UK, the Office for Standards in Education (Ofsted, 2020) has trialed a new methodology for inspection evidence with a focus on ITE partnerships in teacher preparation. In this initiative, the focus is on "how well the centrally taught programme is known and embedded by mentors" in the school context and "how centre-delivered and school-based training have blended to create a coherent experience for trainees" (p. 8). Common within these different examples of schooling and teacher education reform initiatives is the increasing focus on standards, more targeted assessments, and the use of data systems, evidence, and interventions, including through strengthening university and school partnerships and career mentoring for teachers.

The nature and function of standards in teacher education

In these different orientations to reform, the nature and function of standards and teacher expertise are of special interest. Livingston and Flores' (2017) review of trends in teacher education over 40 years in papers published in the *European Journal of Teacher Education* (EJTE) found that "the first paper on teaching standards published in EJTE was in 1982 (Vol. 5, No. 3)" (p. 559). They reported that

> open-mindedness, which is seen as a critical weapon in a teacher's armoury, is endangered by the movement to establish objective standards in education generally and in teacher education in particular … note[ing] that papers that have a specific focus on standards were next published in 2001 and standards have continued to be a significant focus through the 2000s.
>
> *(p. 559)*

In their commentary, Livingston and Flores (2017) characterized standards as putting teacher educators' open-mindedness at risk. This characterization of standards as regulatory and constraining teachers' professionalism and autonomy in practice, is not uncommon in the published research on teacher education (e.g., Beck, 2009; Lambert & Gray, 2020; Lewis et al., 2019). Furlong (2015) identified the significant issue of how standards are conceptualized and how their function is understood. He criticized the use of professional standards in teacher education where they function as "a *de facto* curriculum" (p. 12) which can constrain the design of programs. In such a de facto curriculum approach, the standards can be used as a checklist of unrelated elements to be included in preparation programs and separately assessed and achieved.

In Australia, the TEMAG review also took up the issue of how national professional standards for teachers and program accreditation standards function in ITE. The review reported that these "are weakly applied" (Craven et al., 2014a, p. xi), an observation also reported by the Teaching Council of Aotearoa New Zealand (formally the Education Council of New Zealand, 2016) and by Donaldson (2010) in Scotland, in recommending greater clarity and coherence of standards and expectations within teacher education programs. The TEMAG review recommended that the national professional standards and standards for program accreditation provide "a solid foundation for reform" (Craven et al., 2014a, p. 1).

In this book, we offer a counter-narrative to standards as the de facto curriculum, as characterized by Furlong (2015), with their primary function as regulatory. Instead, we take the view of standards as enabling inquiry and evidence-informed reflection on practice by teacher educators and preservice teachers to promote student learning. Further, we support Cochran-Smith's (2021) stance that accountability is neither inherently positive nor negative. We aim to show an approach to teacher education where teacher educators take the agency to drive reform. Specifically, we aim to show how standards and evidence, taken together, can contribute to teachers' repertoires of foundational knowledge and skills including those necessary to be open-minded, fostering self-regulation in practice.

As presented in the chapters of this book, within this counter-narrative is a sustained program of research and development into culture change in teacher education. (For related discussion of culture change, see Wyatt-Smith et al., 2021). An interdisciplinary team is undertaking this research and includes researchers in assessment and evaluation and teacher education, scholars in data analytics and statistics, digital architects, teacher educators, teacher education regulatory authority personnel, and senior policy officers in professional associations. The actions and decisions, challenges and experiences, and enablers and barriers of attempting such change are the subjects of this book. The account concerns collaboration among multiple stakeholders; building capacity in the use of evidence by preservice teachers and teacher educators; the intrinsic linking of the academic and school-based programs; and finally, the use of digital infrastructure and other resources to sustain teacher education reform.

The move to teaching performance assessments

The chapters present the move to TPAs against a background of three related phenomena, identified previously as issues of 'quality': (1) The rising concern in several countries, including Australia, about how, and how effectively, beginning teachers are prepared for classroom practice; (2) the strengthening interest in professional standards including their potential for leveraging improvement in the quality of the teaching workforce; and (3) the now widespread recognition of the need for assessment capable teachers. As discussed, 'quality' teaching and teachers have been a sustained focus of policy and community expectations for some decades in several countries. However, limited sustained research has been undertaken on methodologies to determine with demonstrated reliability whether a suitable level of 'quality' has been attained for entry to the profession. In Australia, the TEMAG review made a number of recommendations to strengthen the focus on quality assurance of teacher education that included measures on the effectiveness of teacher education programs and their relationship to the practicum. The Australian Government accepted many of the recommendations of the report, including the need for "robust assessment of graduates to ensure classroom readiness" (Australian Government

Department of Education and Training, 2015, p. 2). This resulted in the requirement for a TPA to be completed by all teacher education graduates prior to graduation.

The move to TPAs in ITE is relatively new, representing unchartered territory in most countries. For example, New Zealand is planning to introduce a TPA, with the aim of strengthening ITE (Teaching Council of Aotearoa New Zealand, n.d.). In Australia, responsibility was placed with the Australian Institute for Teaching and School Leadership (AITSL[4]) by the national Government to investigate TPAs as a requirement for ITE program accreditation; to guide teacher education providers regarding how "evidence of their [preservice teachers'] classroom readiness" (Australian Government Department of Education and Training, 2015, p. 8) should be collected; and to work with states and territories to implement these changes. The competence assessment was to be informed by the Australian Professional Standards for Teachers (APST; AITSL, 2011) and the national program standards (AITSL, 2015). Implementation of a TPA is now a requirement "for cohorts completing their program from 2018" (AITSL, 2017, para. 19), that is, for all ITE programs in Australia. A distinguishing feature of the move to TPAs has been the mandatory requirement for moderation as indicated below:

> Program Standard 1 requires that a teaching performance assessment be situated in a classroom environment, to demonstrate a range of teaching practices, and that the assessment is valid, reliable and moderated. In other words, the assessment must:
>
> - Assess the actual practices of teaching and be aligned to the graduate teacher standards
> - Be assessed in a reliable and consistent manner against clear and measurable achievement levels to ensure all pre-service teachers are robustly assessed
> - Include a moderation process to give assurance of the consistency of assessment decisions.
>
> *(AITSL, 2015, p. 10)*

The above makes clear the function of moderation to assure consistency of judgment decisions. However, what counts as effective moderation and the data to show its contribution to the consistency of assessment has received scant official attention. This remains the case some six years after the assurance of consistency through moderation was identified. Readers interested in the relationship between calibration, moderation, and consistency of judgment are referred to the final chapter. Other relevant chapters include Chapter 6 where moderation in standard setting is discussed and Chapter 7 where online cross-institutional moderation (CIM) is considered. While there are significant variations in the term 'moderation' as highlighted in Chapter 7, common across these is the widespread recognition that moderation is part of quality assurance systems and processes and is typically associated with efforts to achieve reliability. In Australia, it has also been associated with efforts to strengthen teachers' assessment capabilities in schooling and higher education. Moderation is an expected practice for classroom teachers (AITSL, 2011) and is a recurring feature of university assessment policy.

The context for our analysis in this book is the establishment and design of a national TPA, called the Graduate Teacher Performance Assessment (GTPA®).[5] The project began in 2016, in the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU; https://www.graduatetpa.com). As key leaders in the project, we established the national Collective of universities. From the beginning, this has involved collaborations and networks of teacher educators, state regulatory authorities, employing authorities, schools, unions, and AITSL, established by the Commonwealth Government to drive reform and accountability for the teaching profession. The development, 2017 trialing, and subsequent implementation of the GTPA have tackled the big questions about ways to connect the necessary knowledge, skills, and theoretical development essential for demonstrating effective teaching, learning, and assessing practice.

The requirement for a TPA has heralded an unprecedented shift in ITE policy in Australia, with the full repercussions yet to be realized. It is already clear that Australian universities are experiencing intense change as they engage in a policy-driven reform of ITE where the focus is sharply on standards and evidence to show graduate competence. These requirements include participation in moderation (see Chapter 7) and arriving at "consistent judgements against the teaching performance assessment rating scales/rubrics" (AITSL, n.d., p. 10). Currently, there are no stated principles for what counts as moderation and no requirement for publicly reporting moderation outcomes. The desirability of a common standard in the country therefore remains unclear and the expected standard for achieving an overall pass on a TPA, shrouded in mystery.

The turn to standards, evidence, and quality through competence assessment

In this book, the GTPA is characterized as an in-the-field catalyst for teacher education reform led by researchers and teacher educators with policy support. It has mobilized a large collective of national partners in teacher education to build an evidence base to show quality in teacher preparation. Current thinking about competence assessment to establish preparedness for teaching is understood as the nexus between the university-based academic program and the school-based practical preparation in the classroom. The linking of these two elements in teaching preparation has long been recognized as both valuable and difficult to achieve (Bloomfield et al., 2013; Donaldson, 2010; Ure et al., 2017). The role of the GTPA as a professionalizing activity, for both teaching and teacher education in Australia, is evident in its development, technical implementation, and the customized approach to CIM, as discussed in this book. The approach to CIM and specifically CIM-Online™,[6] taken in the GTPA Research and Development project, is internationally distinctive (see Wyatt-Smith & Adie, 2021) and has significant implications for efforts to build public confidence in teacher education and in turn, the status of the profession.

In writing the book, we were aware of how teacher education is often characterized as a problem, as previously identified. In response, we decided to take up the invitation from Hutt et al. (2018) to explore the possibility of a proactive role in setting the terms of the debates regarding teacher quality, professionalization, accountability, and innovation in teacher preparation. Our challenge was not to defend ITE but rather to explore how a multidisciplinary team of researchers, teacher educators, data scientists, and digital designers could come together around a newly designed, research-informed assessment of graduate teacher competence as part of a large-scale reform initiative of teacher education in Australia.

The book provides a scholarly examination of large-scale collaboration in a changed standards-driven education policy context. It presents key insights into the professionalization of teaching and intensifying accountability associated with the assessment of teaching graduates on course completion. At issue is the utility of TPAs in establishing graduate preparedness for 'quality teaching'. In addition to addressing this issue, a related intention is to offer critical insights into the conditions in which the teaching profession can exercise agency in policy-driven reform. We propose that this requires teacher educators, researchers, and policy personnel collaborating to inquire into the demanding questions of which standards to apply, what constitutes evidence, and what are the recognizable characteristics for determining teaching competence. We further propose that while these dimensions are essential, they are in themselves insufficient. Issues of methodology, the use of digital technologies, and the collection of large-scale data are additional dimensions that are essential in the move to TPAs as discussed in various chapters.

The book is theoretically framed within assessment as collaborative critical inquiry (Delandshere, 2002; Wyatt-Smith & Gunn, 2009). This involves investigation into complex and interconnected issues of standards, evidence, and impacts of change in diverse contexts of universities, schools, and other sites of teacher preparation. Here, we recognize that universities offer the academic program, and schools act as the sites of professional practice. We also recognize the reported disconnect between the academic program and the practical school-based program, sometimes heard in statements such as 'You'll learn to become a teacher when you start teaching in a school'. Is there a way to make new connections for teachers in preparation so they can avoid the reported theory–practice divide?

One of our starting propositions is that this divide is not only unhelpful but dangerous. How so? It assumes that practice is somehow unrelated to theory, as though practice could ever be other than a demonstration of theory being enacted, though not recognized as such. We also approach the work of teachers from the perspective that teaching, learning, and assessment are influenced by a complex network of expectations, though some of these may well remain unarticulated, but function nevertheless as powerful determinants of what counts as expected practice. This sets the scene for the narrative in this book about how a large national group of teacher educators came together to create change in teacher education, with their own expertise in practice and insights from research and policy to guide the paths taken individually and collectively.

The main objectives of the book are to present: (1) The previously untold account of design decisions, actions, interactions, and relationships that occurred during the development, validation, standard setting, and implementation of the GTPA; (2) new thinking about CIM and benchmarking with accountability vested in the hands of teacher educators; and (3) innovative design and applications of digital architecture for custom-designed apps and secure data storage solutions. Our aim is to offer a coherent theorized and empirically validated response to calls for accountability through:

1. Profiling large-scale collaboration and partnerships that challenged geographic and disciplinary silos in teacher education preparation.
2. Applying judgment and decision-making methodologies that had not previously been combined in standard setting in ITE.
3. Designing and implementing a new approach to CIM-Online™ and benchmarking across multiple teacher education institutions, states, and territories.
4. Designing and implementing longitudinal investigations into candidates' trajectories through teacher preparation and into the teaching workforce.

These four interrelated development lines are part of the larger enterprise of the GTPA research project designed to springboard from a competence assessment as a single instrument to investigate teacher preparation programs and their impact on the learning of their students, post-graduation. The larger enterprise therefore seeks to connect quality preparation for teaching and quality of teaching practice. While the authors recognize that the book is anchored historically in a reform context, our aim is that the discussion of standards and change processes has salience across contexts and time. The theoretical and methodological insights and the treatment of change processes presented in the book are offered as having relevance to those in teacher education and in other professions with an interest in monitoring standards and performance trends over time. Four interlinking themes are at the heart of the book. These are standards, large-scale evidence, professional judgment, and the affordances of digital infrastructure.

Standards. In taking up this theme, the book explores the purposes of standards to gauge the quality of practice demonstrated by teacher education graduates. The approach taken is to use standards and evidence to close the loop that connects standards as inputs in program design, to standards as outputs showing professional competence on program completion. While standards have been available for some time and in various forms internationally (Sachs, 2003; Seameo Innotech Regional Education Program, 2010; Wyatt-Smith & Looney, 2016), these have been used primarily for program design and accreditation purposes. However, there has been no sustained focus on the links across standards, evidence of preparedness to enter the workforce, and the impact of teacher preparation programs on preservice teacher learning and subsequently the learning of their students, post-graduation (for one example of research into teacher employment pathways, see Mayer et al., 2017).

Large-scale evidence and the challenge of building an evidence base. In this second theme, the book explores the actions and decisions taken by a national collective of universities – known as the GTPA Collective – in deciding the evidence needed to 'show' graduate readiness. Thus, the focus is on graduates at the conclusion of the teacher preparation program, rather than the quality of candidates entering teacher preparation. As mentioned earlier, the most recent review of teacher education in Australia (Craven et al., 2014a) identified that the generation of evidence to show the quality and impact of programs on graduates from teacher education was among the highest education policy priorities. In taking up this challenge, a Workforce Studies suite of projects has been designed to examine preservice teacher trajectories over the course of the candidature from entry to exit and into the workforce (see Chapter 11). For the first time in the history of Australia, evidence to show the quality of teaching graduates has been established.

Professional judgment and locating intelligent accountability with the profession. In this theme, the focus is on recursive decisions and actions taken in validation, standard setting, standards-referenced moderation, and benchmarking across institutions. The discussion shows how the GTPA, intended in policy for summative and licensure purposes, is being applied for formative purposes to guide curriculum program planning and review for improvement. These formative purposes go beyond the more narrowly defined and well-recognized summative purposes for TPAs. A distinctive feature of the Australian work presented in this book is the conceptualization, design, and implementation of a dynamic feedback loop that connects professional standards as inputs, to evidence of professional competence as assessed on program completion, as mentioned above. The book explores the decisions that shaped the work and were made in complex networks of diverse groups in ITE and how these have been supported through customized use of digital technologies.

The affordances of the digital infrastructure in sustaining community and to build an evidence base. The criticality of this theme is evident throughout the chapters in efforts to move beyond policy to agentic action by teacher educators. The GTPA project has involved the development and utilization of a system of customized information technology infrastructure that consists of (1) online submission of selected performance samples used for CIM; (2) a web portal for online cross-institutional scoring of samples against the standard; (3) a purpose-designed data app for online collation and storage of data on ITE performance for cohorts of preservice teachers; and (4) data analysis and automated report generation to provide confidential evidence to universities implementing the new competence assessment. Collectively this system, known as the Evidence for Quality in Initial Teacher Education (EQuITE; see Chapter 7), forms a digital architecture designed for the core purpose of providing feedback for curriculum review and program renewal. The data generated through EQuITE are used to build an evidence base on the effectiveness of teacher education, and to improve the quality of teacher education through networks and collaborative partnerships across universities, sector authorities, and government at state and national levels in Australia. The pursuit to build an evidence base to show quality in teacher education is undoubtedly in its infancy in many countries, including in Australia. Given this, the most recent review into teacher education in Australia could be said to be premature, noting that the TEMAG reforms, and, in particular, the introduction of TPAs, have not been bedded down in all universities. More specifically, the core problem, namely that Australia lacks an evidence base to show the quality of teacher education, has not been advanced through government action in any significant way. This book presents an account of the research-led initiative to fill this void in undertaking productive reform in teacher education through large-scale collaboration and digital innovation.

Overview of chapters

The book is presented in two parts. Part 1 (Chapters 1–5) presents the thinking and actions about the conceptualization, design, and implementation of TPAs.

Chapter 1 presents the socio-political context which has provided the genesis to move to reform teacher education. It discusses reviews of teacher education and conditions relevant to the emergence of TPAs. Finally, it introduces four interlinking themes of the book namely standards, large-scale evidence, professional judgment, and the affordances of digital infrastructure.

Chapter 2 presents a discussion on competence assessment. This is considered in selected professions, exploring issues of relevant evidence in the demonstration of competence. The chapter also explores what can be learned from these assessments to inform the development of TPAs in education. Of special interest is the introduction of two TPAs in the United States, both used as policy levers intended to reform teacher education and improve the quality of teaching graduates.

Chapter 3 looks at TPAs through the lens of standards, addressing how standards have informed the design of TPAs. It presents a new conceptualization of TPAs as underpinned by three dynamically interlinked elements – authenticity, system and site validity, and intelligent accountability. We propose that this conceptualization is essential in designing a TPA that is recognized and accepted by the profession. The GTPA is presented as a case instance of how the elements, taken together, constitute underpinning design principles of TPAs.

Chapter 4 focuses on the design features of the GTPA as an authentic assessment of competence for teaching. The chapter addresses the concept of 'readiness' for professional practice and how this may be demonstrated and assessed. It presents a discussion of what preservice teachers need to know and be able to do, extending to consideration of teacherly dispositions. It presents a foundation for subsequent chapters and introduces the core practices of the GTPA.

Chapter 5 focuses on issues of fairness and effectiveness of a TPA used for degree completion and credentialing of future teachers, and for accreditation of ITE programs. Cumming and Pullin highlight educational, technical, and social science issues associated with TPAs in the context of the United States and Australia. Their writing illuminates key aspects in TPA development and implementation, including as they relate to technical standards for assessing professional practice.

Part 2 of the book (Chapters 6–11) carries forward the focus on actions and decision-making, drawing on collective expertise in a range of fields: Teacher education, assessment, standards, data analytics, and systems thinking. Digital architecture and customized design of new infrastructure have been necessary to enable networking at scale. Through a multidisciplinary approach, we discuss building an evidence base to show quality in ITE and promote teacher educator agency in the process.

Chapter 6 presents two methodologies that the researchers chose for validation, standard setting, and establishing reliability of the scoring rubric in the year-long trial of the GTPA. The methods, applied in two separate studies, were the Dominant Profile Judgment Method and Pairwise Comparison. The chapter presents processes for building evaluative expertise resulting in improved reliability of judgments within validation and standard-setting processes. The chapter addresses the challenges in establishing an acceptable standard of performance for entrance to teaching.

In Chapter 7, we argue that demonstrating comparability of standards to determine profession readiness requires principled and rigorous approaches to CIM. These necessarily go beyond the confines of moderation within a single institution in order to generate data showing the valid and reliable application of the standard across institutions. The chapter presents a new approach to benchmarking through online cross-institutional standards-referenced moderation (CIM-Online™). We propose that standards-referenced moderation and benchmarking online are necessary conditions for developing: (1) The dependability and defensibility of teacher educators' judgments and (2) the confidence of teacher educators, preservice teachers, and the public in the quality of graduates entering the teaching profession. We also argue that a combination of principles of fidelity, decision-aids, and calibration training provide complementary means to engage teacher educators in moderation and build dependability of their judgments. Finally, we propose that the model, with its in-built digital architecture and data analytics, is integral to improving program effectiveness and could be applicable in other professions.

In Chapter 8, we assert the need for standards to be part of a feedback loop connecting standards as inputs into teacher education and evidence of standards demonstrated as outputs from teacher education. The focus is not on a mechanistic approach to standards or an atomistic approach to assessing competence. Instead, it is about professional judgment and evaluative expertise needed to discern how the requirements of standards written in qualitative terms can be satisfied in a range of ways. The reform initiative that is the focus of this book infuses professional standards into teacher education in a systematic way through the joint focus on evidence to show graduate competence and the quality and impact of ITE programs. These efforts directly connect standards and evidence, creating a feedback loop in teacher education designed to contribute to building a large-scale evidence base used by teacher educators as a resource for investigating program effectiveness and undertaking review and renewal.

Chapter 9 draws on the core notions of standards, evidence, and accountability in ITE to present accounts of those who have been directly involved in implementing the GTPA. Through applying an ethnomethodological approach to analyzing talk and interactions, the discussion shows how a community of teacher educators navigated system and site requirements to implement a new high-stakes assessment in their respective universities. Against this backdrop, the chapter presents illustrative examples of teacher educators working together through shared inquiry to solve problems, enhance practice, and advance the knowledge base of teacher education.

In Chapter 10, Pullin and Cumming present some of the fairness issues and legal implications of TPAs and the part these play in the licensure of educators. With reference to the United States and Australia, the chapter encompasses technical features of such assessments from a legal perspective. It also considers public policy contexts and non-legal interpretations of fairness issues such as 'opportunity to learn' and processes and procedures in place to address potential injustice of decisions based on the assessment.

Chapter 11 brings together the main elements of the book into an account of teacher education at a watershed in Australia. The authors look back, look sideways, and look forward in our discovery journey into TPAs. In looking back, we reflect on how the story could have developed as an account of government-driven reform of teacher education, externally imposed. This could have been a story of top-down reform, increasing regulation, and anticipated compliance. The alternative telling presents insights into changing cultures in teacher education through collaboration and a sustained focus on professional judgments, standards, and data at scale. This includes the decisions and actions involved in a complex set of partnerships that involved government agencies, universities and schools, unions, and employing authorities. Of interest are the enabling conditions and barriers to mobilizing for change, where the declared intent of the research team and teacher educator Collective is for professional responsibility and accountability. The chapter opens out the vista to ITE workforce studies that examine the trajectories over the period of candidature and follow graduates into the workforce.

Conclusion

As you read this book, we invite you to keep in mind wicked questions concerning (1) the source of real change in a profession, in this case, teacher education, (2) the impact of change on teacher education, and (3) how we can know that it brings benefits to learners and broader society. In the account given in the book, policy and research opened a portal for change. While there was a reported appetite for change, the reality of progressing actual culture change in teacher education is complex and demanding. To achieve culture change in any profession requires time (change does not happen quickly), catalysts for change, and change leaders. It therefore requires engagement at scale, sustainable processes and communication networks, and new mindsets about work and collaboration. This holds true for teacher education and the role it can play in developing quality teachers for contemporary schooling.

Notes

[1] OP is an acronym for overall position and was used in Queensland, Australia, as a tertiary entrance rank to guide selection into universities. It was replaced with the Australian Tertiary Admission Rank system in 2020, in line with other Australian states and territories.

[2] Now the Teaching Council of Aotearoa New Zealand.

[3] *Teacher Education Ministerial Advisory Group Issues Paper*, Creative Commons 'CC BY' 3.0 AU license.

[4] AITSL is a corporate entity, funded by the Australian Government, which has responsibility for establishing Australia-wide teacher professional standards. It has oversight for initial teacher education (ITE) program requirements. It does not register/certify teachers which remains the responsibility of state authorities.

[5] Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

[6] Acknowledgment: The online model of cross-institutional moderation (CIM-Online™) was conceptualized and developed in the Institute for Learning Sciences and Teacher Education, Australian Catholic University. For a discussion of CIM-Online™, readers are advised to also see Wyatt-Smith and Adie (2021).

References

Adams, R. (2021, August 19). Cambridge to end teacher training if government enacts overhaul. *The Guardian*. https://www.theguardian.com/education/2021/aug/18/cambridge-to-cease-teacher-training-if-government-continues-with-damaging-reforms?CMP…

Adie, L. E., Stobart, G., & Cumming, J. J. (2020). The construction of the teacher as expert assessor. *Asia-Pacific Journal of Teacher Education*, *48*(4), 436–453. https://doi.org/10.1080/1359866X.2019.1633623

Advisory Panel for Teacher Education. (2020). *Transforming Norwegian teacher education: The final report of the international advisory panel for primary and lower secondary teacher education*. Norwegian Agency for Quality Assurance in Education. https://www.nokut.no/globalassets/nokut/rapporter/ua/2020/transforming-norwegian-teacher-education-2020.pdf

Australian Government Department of Education and Training. (2015). *Australian government response – Action now: Classroom ready teachers report*. Department of Education, Skills and Employment. https://docs.education.gov.au/documents/australian-government-response-action-now-classroom-ready-teachers-report

Australian Institute for Teaching and School Leadership (AITSL). (2011). *Australian professional standards for teachers*. https://www.aitsl.edu.au/teach/standards

Australian Institute for Teaching and School Leadership (AITSL). (2015). *Accreditation of initial teacher education programs in Australia: Standards and procedures*. https://www.aitsl.edu.au/docs/default-source/default-document-library/accreditation-of-initial-teacher-education-programs-in-australia_jan_2019.pdf?sfvrsn=4639f33c_2

Australian Institute for Teaching and School Leadership (AITSL). (2017). *Fact sheet - Teaching performance assessment - Program Standard 1.2*. https://www.aitsl.edu.au/deliver-ite-programs/teaching-performance-assessment/fact-sheet---teaching-performance-assessment---program-standard-1.2

Australian Institute for Teaching and School Leadership (AITSL). (n.d.). *Teaching performance assessment services: Principles of operation*. https://www.aitsl.edu.au/docs/default-source/initial-teacher-education-resources/tpa/aitsl-tpa-operational-principles_final.pdf?sfvrsn=a50fd3c_2

Baker, J. (2021, February 17). Former premier warns teaching profession facing crisis, change urgently needed. *The Sydney Morning Herald*. https://www.smh.com.au/national/nsw/former-premier-warns-teaching-profession-facing-crisis-change-urgently-needed-20210216-p572xt.html

Bauckham, I., Blake, J., Gill, R., Moore, R., & Twiselton, S. (2021). *Initial teacher training (ITT) market review report*. Department of Education. Gov.UK. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/999621/ITT_market_review_report.pdf

Beck, J. (2009). Appropriating professionalism: Restructuring the official knowledge base of England's 'modernised' teaching profession. *British Journal of Sociology of Education*, *30*(1), 3–14. https://doi.org/10.1080/01425690802514268

Bloomfield, D., Chambers, B., Egan, S., Goulding, J., Reimann, P., Waugh, F., & White, S. (2013). *Authentic assessment in practice settings: A participatory design approach*. https://ltr.edu.au/resources/PP10_1784_Reimann_Report_2013.pdf

Broadfoot, P. (2007). *An introduction to assessment*. Continuum International.

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. *Assessment in Education: Principles, Policy & Practice*, *11*(1), 7–26. http://dx.doi.org/10.1080/0969594042000208976

Businge, C. (2019, April 11). What is the new National Teachers policy? *The New Vision*. https://search.proquest.com/docview/2207186788/fulltext/E966DF2855F406FPQ/72?accountid=8194

Caena, F. (2014). Initial teacher education in Europe: An overview of policy issues. *European Commission Directorate-General for Education and Culture School policy/Erasmus*. ET2020 Working Group on Schools Policy. http://ec.europa.eu/assets/eac/education/experts-groups/2014-2015/school/initial-teacher-education_en.pdf

Carey, A. (2021, April 27). Minister says quality teaching, not more school funding key to better results. *The Age*. https://www.theage.com.au/national/minister-says-quality-teaching-not-more-school-funding-key-to-better-results-20210426-p57mfl.html

Carter, A. (2015). *Carter review of initial teacher training (ITT)*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/399957/Carter_Review.pdf

Clark, E. (2017, November 29). Teacher training, pay and performance: What makes a difference to kids? *Australian Broadcasting Corporation*. https://www.abc.net.au/news/2017-11-30/teacher-training,-pay-and-performance-what-gets-results/9187926

Clarke, M., & Parker, K. (2021, August 18). Major teaching reform in England will erode the intellectual basis of the profession. *The Conversation*. https://theconversation.com/major-teaching-reform-in-england-will-erode-the-intellectual-basis-of-the-profession-165102

Cochran-Smith, M. (2020). Relocating teacher preparation to new Graduate Schools of Education. *The New Educator*, *17*(1), 1–20 https://doi.org/10.1080/1547688X.2020.1814466

Cochran-Smith, M. (2021). Rethinking teacher education: The trouble with accountability. *Oxford Review of Education*, *47*(1), 8–24. https://doi.org/10.1080/03054985.2020.1842181

Cochran-Smith, M., Piazza, P., & Power, C. (2013). The politics of accountability: Assessing teacher education in the United States. *The Educational Forum*, 77(1), 6–27. https://doi.org/10.1080/00131725.2013.739015

Collins, S. (2017, December 8). Teacher training failing. *The New Zealand Herald*. https://search.proquest.com/docview/1973448445/1220EB5B60DC461DPQ/83?accountid=8194

Craven, G., Beswick, K., Fleming, J., Fletcher, T., Green, M., Jensen, B., Leinonen, E., & Rickards, F. (2014a). *Action now: Classroom ready teachers*. Teacher Education Ministerial Advisory Group, TEMAG. Department of Education, Australia. https://docs.education.gov.au/documents/action-now-classroom-ready-teachers-report

Craven, G., Beswick, K., Fleming, J., Fletcher, T., Green, M., Jensen, B., Leinonen, E., & Rickards, F. (2014b). *Teacher Education Ministerial Advisory Group Issues Paper*. file:///Users/leadie/Desktop/temag_issues_paper_-_april_2014.pdf

Crowe, E. (2011). *Race to the Top and teacher preparation: Analyzing state strategies for ensuring real accountability and fostering program innovation*. Centre for American Progress. https://www.americanprogress.org/issues/education-k-12/reports/2011/03/01/9329/race-to-the-top-and-teacher-preparation/

Delandshere, G. (2002). Assessment as inquiry. *Teachers College Record*, *104*(7), 1461–1484. https://doi.org/10.1111/1467-9620.00210

Denholm, A. (2017, May 18). Huge gap in colleges' hours of literacy and numeracy. *The Herald*. https://search.proquest.com/docview/1899672458/209197E458074801PQ/1?accountid=8194

Dick, S. (2021, March 11). Australian students suffering poor-quality teaching and testing: Tudge. *The New Daily*. https://thenewdaily.com.au/news/national/2021/03/11/education-australia-oecd-teachers/

Donaldson, G. (2010). *Teaching Scotland's future: Report of a review of teacher education in Scotland*. https://www2.gov.scot/resource/doc/337626/0110852.pdf

Drew, A. (2018, March 28). Teacher training is in need of a radical rethink. *The Herald*. https://search.proquest.com/docview/2018924400/fulltext/D4CD678541304217PQ/137?accountid=8194

Education Council of New Zealand. (2016). *Strategic options for developing future orientated initial teacher education*. https://teachingcouncil.nz/sites/default/files/Strategic%20options%20REVISED%2029%20JUNEpdf.pdf

Ell, F., Simpson, A., Mayer, D., McLean Davies, L., Clinton, J., & Dawson, G. (2019). Conceptualising the impact of initial teacher education. *Australian Educational Researcher*, *46*(1), 177–200. https://doi.org/10.1007/s13384-018-0294-7

European Commission. (2013). *Supporting teacher competence development for better learning outcomes*. https://ec.europa.eu/assets/eac/education/policy/school/doc/teachercomp_en.pdf

Furlong, J. (2015). *Teaching tomorrow's teachers: Options for the future of initial teacher education in Wales*. https://gov.wales/sites/default/files/publications/2018-03/teaching-tomorrow%E2%80%99s-teachers.pdf

García, E., & Weiss, E. (2019). *The teacher shortage is real, large and growing, and worse than we thought*. The Perfect Storm in the Teacher Labor Market Series. epi.org/163651

Gonski, D., Arcus, T., Boston, K., Gould, V., Johnson, W., O'Brien, L., Perry, L., & Roberts, M. (2018). *Through growth to achievement: Report of the review to achieve educational excellence in Australian schools*. https://docs.education.gov.au/system/files/doc/other/662684_tgta_accessible_final_0.pdf

Goss, P. (2017). *Towards an adaptive education system in Australia*. Grattan Institute. https://grattan.edu.au/report/towards-an-adaptive-education-system-in-australia/

Goss, P., Sonnemann, J., & Griffiths, K. (2017). *Engaging students: Creating classrooms that improve learning*. Grattan Institute. https://grattan.edu.au/wp-content/uploads/2017/02/Engaging-students-creating-classrooms-that-improve-learning.pdf

Goss, P., Sonnemann, J., & Nolan, J. (2019). *Attracting high achievers to teaching*. Grattan Institute. https://grattan.edu.au/report/attracting-high-achievers-to-teaching/

Grant, G. (2017, September 1). Three Rs' shame of Scotland's teachers: MSPs call for probe into soaring number of trainees who can't read or write properly: Education decline 'fuelled by skills gap in trainee teachers'. *Daily Mail*. https://search.proquest.com/docview/1934136081/fulltext/2E3B43FC1DFA4412PQ/174?accountid=8194

Green, C., Eady, M., & Andersen, P. (2018). Preparing quality teachers: Bridging the gap between tertiary experiences and classroom realities. *Teaching and Learning Inquiry: The ISSOTL Journal*, *6*(1), 104–125. https://doi.org/10.20343/teachlearninqu.6.1.10

Hagger, H., & McIntyre, D. (2006). *Learning teaching from teachers. Realizing the potential of school-based teacher education*. Open University Press.

Hare, J. (2021a, April 12). How Tudge can get students back to the top of OECD rankings. *Financial Review*. https://www.afr.com/work-and-careers/education/how-tudge-can-get-kids-back-to-the-top-of-the-class-20210411-p57i8m

Hare, J. (2021b, April 15). New review into how to attract the best and brightest into teaching. *Financial Review*. https://www.afr.com/work-and-careers/education/new-review-into-how-to-attract-the-best-and-brightest-into-teaching-20210414-p57j4u

Hattie, J. (2003, October 19–21). *Teachers make a difference: What is the research evidence?* [Conference Paper]. Australian Council for Educational Research Annual Conference, Melbourne, Australia. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1003&context=research_conference_2003

Haugsbakk, G. (2013). From Sputnik to PISA shock: New technology and educational reform in Norway and Sweden. *Education Inquiry*, *4*(4), 607–628. https://doi.org/10.3402/edui.v4i4.23222

Henebery, B. (2020, January 16). Schools face 'critical' teacher shortage in 2020. *The Educator*. https://www.theeducatoronline.com/k12/news/schools-face-critical-teacher-shortage-in-2020/269422

Hutt, E. L., Gottlieb, J., & Cohen, J. J. (2018). Diffusion in a vacuum: edTPA, legitimacy, and the rhetoric of teacher professionalization. *Teaching and Teacher Education*, *69*, 52–61. https://doi.org/10.1016/j.tate.2017.09.014

Johnson, S. (2017, May 18). Trainee teachers spend as little as four hours on literacy and numeracy. *The Daily Telegraph*. https://search.proquest.com/docview/1899659322/fulltext/1220EB5B60DC461DPQ/41?accountid=8194

Jones, N. (2017, April 11). Time to raise bar? *The New Zealand Herald*. https://search.proquest.com/docview/1885912571/6728272795BD46DFPQ/152?accountid=8194

Lambert, K., & Gray, C. (2020). Performing 'teacher': Exploring early career teachers' becomings, work identities and the [mis-]use of the professional standards in competitive educational assemblages. *Pedagogy, Culture & Society*, *28*(4), 501–523. https://doi.org/10.1080/14681366.2019.1663247

Lewis, S. C., Savage, G., & Holloway, J. (2019). Standards without standardisation? Assembling standards-based reforms in Australian and US schooling. *Journal of Education Policy*, *35*(6), 737–764. https://doi.org/10.1080/02680939.2019.1636140

Lingard, B., Wyatt-Smith, C., & Heck, E. (2021). Transforming schooling through digital disruption. In C. Wyatt-Smith, B. Lingard, & E. Heck (Eds.), *Digital disruption in teaching and testing: Assessments, big data, and the transformation of schooling* (pp. 1–33). Routledge.

Livingston, K., & Flores, M. A. (2017). Trends in teacher education: A review of papers published in the European Journal of Teacher Education over 40 years. *European Journal of Teacher Education*, *40*(5), 551–560. https://doi.org/10.1080/02619768.2017.1387970

Louden, W. (2008). 101 Damnations: The persistence of criticism and the absence of evidence about teacher education in Australia. *Teachers and Teaching: Theory and Practice*, *14*(4), 357–368. https://doi.org/10.1080/13540600802037777

Mayer, D., Dixon, M., Kline, J., Kostogriz, A., Moss, J., Rowan, L., Walker-Gibbs, B., & White, S. (2017). *Studying the effectiveness of teacher education: Early career teachers in diverse settings*. Springer.

McKinsey & Co. (2007). *How the world's best-performing school systems come out on top*. https://www.mckinsey.com/~/media/mckinsey/industries/public%20and%20social%20sector/our%20insights/how%20the%20worlds%20best%20performing%20school%20systems%20come%20out%20on%20top/how_the_world_s_best-performing_school_systems_come_out_on_top.pdf

Moir, J. (2017, December 8). New teachers 'ill-equipped for job'. *The Press*. https://search.proquest.com/docview/1973502731/fulltext/D801D0B8FE694321PQ/102?accountid=8194

O'Flaherty, A. (2020, January 16). Low OP hurdle dumbing down future teachers. *Courier Mail*, 15.

Office for Standards in Education (Ofsted). (2020). *Consultation proposals for the framework to inspect the quality of teacher education from September 2020*. https://www.gov.uk/government/consultations/initial-teacher-education-inspection-framework-and-handbook-2020-inspecting-the-quality-of-teacher-education/consultation-proposals-for-the-framework-to-inspect-the-quality-of-teacher-education-from-september-2020

Patty, A. (2021, January 17). Teacher shortage opens gate to country lifestyle. *The Sydney Morning Herald*. https://www.smh.com.au/business/workplace/teacher-shortage-opens-gate-to-country-lifestyle-20210111-p56t86.html

Pyne, C. (2014, February 18). A quality education begins with the best teachers, says Christopher Pyne. *The Sydney Morning Herald*. https://www.smh.com.au/politics/federal/a-quality-education-begins-with-the-best-teachers-says-christopher-pyne-20140219-32z61.html

Rauschenberger, E., Adams, P., & Kennedy, A. (2017). *Measuring quality in ITE literature review: A Literature Review for Scotland's MQuITE Study*. http://www.scde.ac.uk/wp-content/uploads/2017/10/MQuITE-Lit-Review-FINAL-Oct-2017.pdf

Rickenbrode, R., Drake, G., Pomerance, L., & Walsh, K. (2018). *2018 Teacher Prep Review (NCTQ)*. https://www.nctq.org/publications/2018-Teacher-Prep-Review

Robinson, N. (2019, January 10). Managing teaching standards: Why South Africa will find it hard to break free from its 'vicious' teaching cycle. *The Daily News*. https://search.proquest.com/docview/2165541425/fulltext/66D2F3D6D3D547EEPQ/25?accountid=8194

Sachs, J. (2003). Teacher professional standards: Controlling or developing teaching? *Teachers and Teaching*, *9*(2), 175–186. https://doi.org/10.1080/13540600309373

Sahlberg, P. (2012). *Report of the international review panel on the structure of initial teacher education provision in Ireland: Review conducted on behalf of the Department of Education and Skills*. https://www.education.ie/en/Publications/Education-Reports/Report-of-the-International-Review-Panel-on-the-Structure-of-Initial-Teacher-Education-Provision-in-Ireland.pdf

Sahlberg, P., Broadfoot, P., Coolahan, J., Furlong, J., & Kirk, G. (2014). *Aspiring to excellence - final review (Nth Ireland)*. https://dera.ioe.ac.uk/20454/1/aspiring-to-excellence-review-panel-final-report.pdf

Schwab, K. (2017). *The fourth industrial revolution*. Portfolio Penguin.

Seameo Innotech Regional Education Program. (2010). *Teaching Competency Standards in Southeast Asian Countries*. http://www.seameo.org/_files/SEAMEO_Teaching_Competency_Standards-WTD.pdf

See, B. H., & Gorard, S. (2020). Why don't we have enough teachers?: A reconsideration of the available evidence. *Research Papers in Education*, *35*(40), 416–442. https://doi.org/10.1080/02671522.2019.1568535

Targeted News Service. (2018, October 5). UNICEF issues joint statement on World Teacher's Day. *Targeted News Service*. https://search.proquest.com/docview/2116594401/fulltext/1220EB5B60DC461DPQ/75?accountid=8194

Teaching Council of Aotearoa New Zealand. (n.d.). *Initial teacher education 2021*. https://teachingcouncil.nz/sites/default/files/ITE%20detail%20decisions%20and%20vision.pdf

The Southland Times. (2017, December 8). New teachers not making the grade. *The Southland Times*. https://search.proquest.com/docview/1973501455?accountid=8194

Tudge, A. (2021, March 11). *Being our best: Returning Australia to the top group of education nations* [Press Release]. https://ministers.dese.gov.au/tudge/being-our-best-returning-australia-top-group-education-nations

U.S. Department of Education. (2016). *Race to the top fund*. https://www2.ed.gov/programs/racetothetop/index.html

UNESCO Institute for Statistics. (2016). *The world needs almost 69 million new teachers to reach the 2030 education goals*. https://unesdoc.unesco.org/ark:/48223/pf0000246124

Ure, C., Hay, I., Ledger, S., Morrison, C., Sweeney, T.-A., & Szandura, A. (2017). *Professional experience in initial teacher education: A review of current practices in Australian ITE*. Australian Government Department of Education and Training.

Visentin, L. (2021, April 15). Teacher training review key to arresting declining academic results: Tudge. *The Age*. https://www.theage.com.au/politics/federal/teacher-training-review-key-to-arresting-declining-academic-results-tudge-20210414-p57j6i.html?fbclid…

Volante, L., DeLuca, C., Adie, L., Baker, E., Harju-Luukkainen, H., Heritage, M., … Wyatt-Smith, C. (2020). Synergy and tension between large-scale and classroom assessment: International trends. *Educational Measurement: Issues and Practice*, *39*(4), 21–29. https://doi.org/10.1111/emip.12382

Waldow, F. (2009). What PISA did and did not do: Germany after the "PISA-shock." *European Educational Research Journal*, *8*(3), 476–483. https://doi.org/10.2304/eerj.2009.8.3.476

Wanzala, O. (2019, February 9). Teacher training set for major changes. *Daily Nation*. https://search.proquest.com/docview/2177373303/fulltext/2E3B43FC1DFA4412PQ/214?accountid=8194

Waugh, A. (2020, January 24). Cardy: Universities have to improve teacher training. *Postmedia Network*. https://search.proquest.com/canadiannews/docview/2344085797/fulltext/26C19FD8E40A4527PQ/1?accountid=8194

Wiggan, G., Smith, D., & Watson-Vandiver, M. J. (2020). The national teacher shortage, urban education and the cognitive sociology of labor. *Urban Review*, *53*(1), 43–75. https://doi.org/10.1007/s11256-020-00565-z

Wightwick, A. (2017, June 15). New training board members announced. *Western Mail*. https://search.proquest.com/docview/1909948809/1220EB5B60DC461DPQ/40?accountid=8194

Wyatt-Smith, C., & Adie, L. (2021). Introducing a new model for online cross-institutional moderation. In C. Wyatt-Smith, L. Adie, & J. Nuttall (Eds.), *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration* (pp. 39–58). Springer.

Wyatt-Smith, C., Adie, L., & Nuttall, J. (Eds.). (2021). *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration*. Springer.

Wyatt-Smith, C., Alexander, C., Fishburn, D., & McMahon, P. (2017). Standards of practice to standards of evidence: Developing assessment capable teachers. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 250–270. https://doi.org/10.1080/0969594X.2016.1228603

Wyatt-Smith, C., & Gunn, S. (2009). Towards theorising assessment as critical inquiry. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 83–102). Springer.

Wyatt-Smith, C., & Looney, A. (2016). Professional standards and the assessment work of teachers. In D. Wise, L. Hayward, & J. Pandya (Eds.), *The SAGE handbook of curriculum, pedagogy and assessment* (pp. 805–820). Sage.

Yeigh, T., & Lynch, D. (2017). Reforming initial teacher education: A call for innovation. *Australian Journal of Teacher Education*, *42*(12), 112–127. https://doi.org/10.14221/ajte.2017v42n12.7

2

THE MOVE TO ASSESSING COMPETENCE IN PROFESSIONS

Introduction

The starting proposition in this book is that successful registration and entry into a profession relies on the quality of pre-registration education and assessment (preparation during candidature) and the quality of assessment on program completion to determine readiness to enter the profession. This observation applies equally well to all professions. The assessment of professional competence for credentialing is already well accepted as a mandatory component of program accreditation, certification, and licensure in some professions. This chapter considers competence assessments in medicine, dentistry, law, psychology, and education to examine their use in determining preparedness for professional practice.

In the field of education, our attention turns to two teaching performance assessments (TPAs) in the United States, namely the Performance Assessment for California Teachers (PACT), and the Educative Teacher Performance Assessment (edTPA). These have been useful reference points in developing TPAs in Australia, a country which came to teaching competence assessment in 2014 (see Craven et al., 2014). The lessons learned from the U.S. TPA experience were salient, given the official expectation that TPAs in Australia were among the proposed policy levers for reforming teacher education and improving the quality of teaching graduates. TPAs were expected to improve readiness for professional practice. The U.S. examples were also relevant for what they could reveal about issues of validity, reliability, and standard setting; fidelity and fairness in test implementation; methodologies for analyzing preservice teachers' submitted performance data; and implications, including legal implications and risks, of the move to TPAs. Readers interested in fairness and legal issues that might arise when students feel that their specific needs are not met in TPA implementation are referred to Chapters 5 and 10, respectively.

DOI: 10.4324/9780429318504-3

Finally, in this chapter, we examine the concept of readiness for the classroom as this has emerged as a force in Australian teacher education reform.

Measuring complex assessments of professional competence

There are multiple, sometimes competing definitions of the term 'competence' and different methodologies for assessing it (Blömeke et al., 2015). In this book, we work from the position that competence is understood to be "a complex combination of knowledge, skills, understanding, values, attitudes and desire which lead to effective, embodied human action in the world, in a particular domain" (Hoskins & Deakin Crick, 2010, p. 121). In the psychology profession, competence has been conceptualized by Gonsalvez and Crow (2014, p. 177) to include the human dimension of professional work as "the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice" (drawing on Epstein & Hundert, 2002). Further, drawing on Falender and Shafranske (2007), Gonsalvez and Crow defined competencies as the "measurable human capabilities involving knowledge, skills, and values, which are assembled in work performance" (p. 177). Blömeke et al. (2015) defined competence as "the latent cognitive and affective-motivational underpinning of domain-specific performance in varying situations" (p. 3). The authors recognized that the cognitive and affective-motivational traits can only "be inferred from observable behavior" (p. 3). They further identified the complexity of measuring with validity and reliability complex performance tasks, raising concerns about inherent measurement error.

Typically, measuring competence involves providing valid and reliable evidence of the latent traits assessed against criteria or observed as performance in context. Competence is understood "to involve a multitude of cognitive abilities and affect-motivation states that are ever changing throughout the duration of the performance" (Blömeke et al., 2015, p. 4) such that "no two people might use the exact same competence profile to carry out the behaviour" (p. 4). This stance informs our first observation that there can be different configurations of the properties or characteristics of a behavior or performance (sometimes referred to as criteria). They can combine in a range of ways subject to influences of context. In so doing, they can represent acceptable variations in the competence profile and offer potentially different ways of satisfying the quality requirements of the expected performance.

However, Gonsalvez and Crow (2014) identified little evidence of the predictive and construct validity in the frameworks developed to conceptualize and organize the valued competencies in the psychology profession. They highlighted the necessity to "determine the dimensional structure and clustering of these many competencies and to chart the normative developmental trajectories for the diverse dimensions" (p. 186). They also acknowledged the difficulty in evaluating the attitude-value competencies. The concern raised by these authors relates to the reduction of performance to knowledge and skills that can ignore context, and the emotional or affective work of the practitioner, a result that is often precipitated by the difficulty with scoring complex performances in authentic contexts.

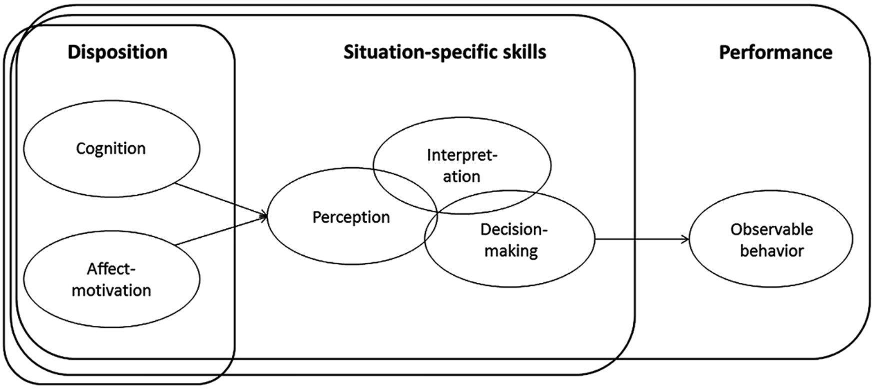


FIGURE 2.1 Blömeke et al.'s model of competence linking disposition and performance through situation-specific skills of reasoning (reproduced with permission from Blömeke et al., 2015, p. 7).

Blömeke et al. (2015) provided a model that incorporates the generic cognitive skills and affective dispositions recognizable within a profession, situated within the perception, interpretation, and decision-making of a particular case instance and related to real-world performance and behavior (Figure 2.1). The details and combinations of skills and dispositions that produce a performance may not readily lend themselves to scrutiny, even by the person delivering the performance; an observation widely reported in the literature on expertise (Dreyfus & Dreyfus, 2005; Feltovich, et al., 2006; Popham, 2017; Sadler, 1985). Blömeke et al. (2015) asserted that by presenting the processes of reasoning, dispositions can be linked to performance in an authentic assessment of competence. This position informs our second observation that such linking of dispositions to performance by making reasoning apparent will be highly relevant to professions that rely on interactions between members of a profession and those they serve in the public good, for example, health professionals, lawyers, teachers, and psychologists.

To this point, we have made two observations: (1) There can be different configurations of characteristics of performance and (2) dispositions can be linked to performance through reasoning processes. Taken together, these suggest challenges associated with assessing performance in the field of practice: Competence is inherently complex and not readily able to be standardized and expectations may not be able to be wholly prescribed and anticipated in advance of the demonstration. These observations apply to competence assessment requiring a demonstration of practice involving talk, actions, interactions, and an account of reasoning and decision-making that shaped the practice in context.

Performance assessments

As mentioned, assessments of professional competence related to licensure for practice are features of professions such as medicine, psychology, and law. Of interest in this book concerning teacher education, and where the Australian research on TPAs started, is what can be learned from the experiences of various professions regarding the use and value of competence assessments of actual practice undertaken in situ. In this chapter, such competence assessments of practice are understood to be performance assessments. The *Standards for educational and psychological testing* (American Educational Research Association/American Psychological Association/National Council on Measurement in Education [AERA/APA/NCME], 2014) identify critical features that distinguish performance assessments and performance standards:

> *Performance assessments* – "assessments for which the test taker actually demonstrates the skills the test is intended to measure by doing tasks that require those skills"
>
> *(p. 221)*
>
> *Performance level descriptors* – "descriptions of what test takers know and can do at specific performance levels"
>
> *(p. 221)*
>
> *Performance standards* – "descriptions of levels of knowledge and skill acquisition contained in content standards as articulated through performance-level labels (e.g., 'basic,' 'proficient,' 'advanced'); statements of what test takers at different performance levels know and can do; and cut scores or ranges of scores on the scale of an assessment that differentiate levels of performance"
>
> *(p. 221)*

When performance assessments are embedded wholly in context and involve actual people and social interactions in real time (e.g., patients, clinicians, and trainees in a dental clinic or hospital), they are distinguishable from simulations of practice completed in a laboratory (Heap, 1987) or from responses to scripted talk and interactions that are video recorded in practice contexts (Maude et al., 2021). Assessments that use videos of actors in scripted interactions in context are described in German research in the preparation of teachers in biology (Kramer et al., 2020) and mathematics pedagogy (König et al., 2021). Preservice teachers watch the videos and consider the behaviors demonstrated in teaching and interactions with students.

An example of a simulation of clinical practice that follows the problem-based learning approach, uses simulation manikins on which the student practices and demonstrates key skills. Such simulations with manikins are typically used in hairdressing, optometry, physiotherapy, and medicine, and generally occur at a stage prior to interacting with actual clients and patients. Problem-based learning and practice approaches, including the use of manikins, are part of the preparation of students in the dentistry program, University of Hong Kong, in which they progress from the manikins to patients in a clinical setting to undertake a set of key skills tests. This move occurs after the students have demonstrated necessary knowledge and skills in a suite of tests working with manikins. An example of an authentic performance assessment is one of the mandatory key skills assessments in the University of Hong Kong dentistry program (Bridges et al., 2017). This requires the student to undertake supervised work with an actual patient in a clinic that is open to the public. The key skills assessment in the clinic has the characteristics of a complex authentic performance assessment, that is, it occurs in a recognizably authentic professional practice setting and involves case-review and decision-making in response to a presenting patient with a real problem requiring diagnosis and action associated with the targeted demonstration of the skill or skills to be assessed.

One of the recognized challenges in implementing performance assessments relates to the need for embedding the assessments in actual contexts, which in turn involves action and decision-making, talk, and other interactions. Put simply, the challenge is how to achieve rater reliability and in turn, achieve public confidence in the measure applied (the instrument) and the measurement (score or rating) that it produces. While the requirements of performance assessments can vary widely within and across professions, a common feature is that they are inherently contextual and can involve responses to multiple variables at any given time. This means that the interpretation of the requirements and how expectations of performance are understood and applied can vary. The authors of this book propose that this phenomenon reflects that performance assessment is necessarily bound up with context and how individuals interact in context. For example, the performance of the student undertaking the assessment in the dental clinic or hospital can reasonably be influenced by patient responses and the role of the supervisor during the performance.

Human scoring of performance assessment is acknowledged as a complex process impacted by a range of variables (Cooksey et al., 2007; Hammond, 1996; Moss, 1994). It is not surprising that reliability of judgment, and the enabling conditions that support high reliability, have been of ongoing interest in educational measurement and assessment literature for decades dating back to Diederich (1974) and has been a continuing feature in the work of assessment scholars (e.g., Harlen, 2005; Messick, 1995). The reliability of an assessment is related to the consistency of scoring and comparability of the judgments (Moss, 1994). It has been described as "a necessary condition for ensuring high measurement quality, but… not a sufficient one" (Newton & Shaw, 2014, p. 14). Several authors including Broadfoot (2010) and Moss (1994) warned of valuing reliability over validity of an assessment which may deter assessment of complex learning outcomes as distinct from discrete, easily measured competencies and skills, for example, in a checklist or in multiple-choice items.

Interest in the reliability and validity of performance assessment is, in part, related to the belief that assessment produces results that can be generalized to a more global domain of interest. Wiliam (1994) stated that "if we cannot have faith in the assessment to tell us anything about what was actually assessed, we can have even less faith in generalizations to wider domains" (p. 18). He argued for connecting reliability and validity to the concept of dependability. The notion of faith in an assessment – its dependability to measure what it claims to measure – is directly tied to public confidence in those awarded licensure to practice. In the medical profession, licensing exams are used in many countries to ensure the conduct of safe medical practice, thus reassuring the public of the competence of practicing doctors.

Archer et al. (2016) conducted a systematic review of the medical literature to establish evidence of the validity of a range of large-scale licensing examinations used as entry requirements before commencing clinical practice. Information on validity evidence was compiled from medical regulators and licensing authorities in 49 countries that had been classified as "very high human development countries similar to the UK" (p. 2 and citing Malik, 2014). Using a validity framework developed by the AERA, the APA, and the NCME (see Downing, 2003), the review did not establish causal evidence between performance on the licensure exams and patient outcomes. Archer et al. (2016) concluded that "the debate on licensure examinations is characterized by strong opinions but is weak in terms of validity evidence" (p. 9) such that the debate continues regarding the value of such examinations in the medical profession. The collation of validity evidence for performance assessments and how this evidence links to competent performance in practice is a key consideration in the introduction of such assessments into teaching.

The development of a performance assessment requires a rigorous approach to establishing the dependability of the instrument, using a process that maintains validity while achieving a high level of reliability against a performance standard (Harlen, 2005). However, as suggested, the judgment of performances can be influenced by different purposes and contexts of assessment, especially where the site of performance can influence the nature and scope of the performance and how the assessment is regarded by others who are influential in the context. Further, the reliability of judgments can vary depending on (1) the interrelationships between the local (site) and system (regulatory) contexts in which the assessment is undertaken and (2) the experiences of the judges in recognizing the performance standard in actual practice and when this has been achieved (Cooksey et al., 2007; Freebody & Wyatt-Smith, 2004; for a more detailed discussion see Chapter 3). Identifying the conditions that promote reliability of judgments in complex performance assessment against a common standard for licensure is widely recognized to be challenging, though essential, when establishing professional competence. It is also critical in building public confidence in the credibility and fairness of the assessment.

Examples of performance assessments for determining professional competence and licensure

The following discussion highlights a reimagining of competence assessment underway in several professions including nursing, paramedicine, psychology, and law. As mentioned earlier in this book, the authors wanted to 'see' competence as it is understood and assessed in other professions so that lessons could be derived to inform the development of performance assessments in the field of teaching in Australia. The scan brought to light significant differences, especially in how competence is assessed, and four recurring characteristics:

1. A key role of professional standards administered by a registration authority or professional association.

2. A growing interest in at least one component of the competence assessment occurring in-the-field.
3. The connection of the theoretical program and the practical component in professional preparation, with an assessment of competence informed by both strands.
4. Oversight of the competence assessment by a professional body – the association responsible for administering registration.

Noting the importance of licensing assessments to guarantee safe medical practice, nursing is one example of "a regulated profession that requires each nurse to meet requisite standards" (Takashima et al., 2019, p. 502). In Australia, nursing students undertake the *Australian Nursing Standards Assessment Tool* (ANSAT; 2014, revised 2018) as part of their mid- (formative) and end clinical placements (summative). The ANSAT is a one-page work-based tool of 23 items: "Each item is scored on a numerical scale from one (not able to perform) to five (performed at an excellent standard) with a three being the passing level of performance that would be expected for the individual student's level" (Takashima et al., 2019, p. 504). The test utilizes the Nursing and Midwifery Board of Australia (NMBA) registered nurse standards: (1) Thinks critically and analyzes nursing practice, (2) engages in therapeutic and professional relationships, (3) maintains the capability for practice, (4) comprehensively conducts assessments, (5) develops a plan for nursing practice, (6) provides safe, appropriate, and responsive quality nursing practice, and (7) evaluates outcomes to inform nursing practice (Australian Health Practitioner Regulation Agency [AHPRA]; Nursing and Midwifery Board AHPRA, 2021). Importantly, as Takashima et al. (2019) commented, "The ANSAT is designed as a workplace appraisal instrument that reports continuing performance of all areas of professional practice rather than a one-off or a staged demonstration, that is often associated with competency assessments" (p. 504). The ANSAT assessment has been noted as a robust and valid assessment tool, with further research being undertaken into the examination of inter-rater reliability (Ossenberg et al., 2020).

Similarly, emerging in the field of paramedicine in Australia, clinical placements are assessed with the AHPRA *Paramedic Competency Assessment.* The AHPRA Paramedicine Board of Australia is responsible for the assessment and the setting of standards for practicing paramedics in Australia. The AHPRA Paramedic Competency Assessment comprises of a two-page assessment tool that is based on the professional capabilities for registered paramedics from the Paramedicine Board (AHPRA) across five domains: (1) The professional and ethical practitioner, (2) the communicator and collaborator, (3) the evidence-based practitioner, (4) the safety and risk management practitioner, and (5) the paramedicine practitioner (Paramedicine Board of Australia, 2021). In addition to this, "The recent launch of national registration for paramedics in Australia coincided with the publication of a set of professional capabilities, setting out the minimum expectations of knowledge and skills for practice under the paramedic title" (Smith et al., 2020, p. 2). This recent publication has led to work around another new assessment, the *Australasian*

Paramedic Competency Assessment Tool (APCAT), designed to assess the practice competency of undergraduate paramedic students (Smith et al., 2020). Currently, the *AHPRA Paramedic Competency Assessment* is being used in the assessment of some applicants (Paramedic Competency Assessment Consortium, n.d.), with five universities forming a consortium in 2019 to conduct competency assessments by the Paramedicine Board. In both nursing and paramedicine, the introduction of professional standards has provided a catalyst for the development of common competency assessments for profession-ready graduates.

In the psychology profession in Australia, entry is guided by the conditions stipulated in the Psychology Board of Australia's *Registration Standard: Provisional Registration* (Psychology Board of Australia, 2017), *Registration Standard for General Registration* (Psychology Board of Australia, 2016), and the Health Practitioner Regulation National Law Act 2009 (Cth). Underpinned by eight core competencies of psychology practice, conditions for registration entail satisfactory completion of a six-year program of education and practice (Psychology Board AHPRA, 2020). This can be completed through a four-year accredited psychology program, after which candidates apply for provisional registration before being able to commence the final two years of their qualification. These two years can be completed through one of three pathways: (1) A higher degree qualification (e.g., Masters or PhD), (2) a five-year degree and one-year internship, or (3) a two-year internship[1]. In the second and third pathways, candidates must pass the National Psychology Examination – a culminating, summative assessment designed to test provisional psychologists' knowledge of psychological practice by examining authentic case studies and related issues encountered in professional practice. The Board-approved 1–2-year internship program comprises 1500–3000 hours of supervised practice, the submission of case reports and bi-annual supervisory reports, and a final assessment of competence report to the Board. Provisional psychologists must achieve an overall score of 70% to demonstrate the "minimum level of applied knowledge required for independent psychology practice" (Psychology Board of Australia, 2019, p. 7). Performances and pass rates are moderated by the regulating authority. All three pathways require assessment of knowledge and skills through theory and performance.

In the legal profession, assessment reform is underway in the UK with the introduction of the Solicitors Qualifying Examination and a period of qualifying work experience (see Bone & Maharg, 2019, p. 3). Maharg and Webb (2019) provided an overview "of legal education reform movements currently taking place in the Common Law world" (p. 25) and discussed international review and innovation in assessment of legal education and practice. Based on this overview and other published information, internationally, there appears to be no reported common or large-scale authentic complex performance assessment for law. For example, in Australia, both national and state-based legislative guidelines inform admission to the legal profession. However, the *Competency Standards for Entry-Level Lawyers* (Law Admissions Consultative Committee, 2015) which cover the skills, practice areas, and values required of a lawyer, must be evident and met within the practical component of legal training. Thus, admission to practice as a lawyer requires satisfactory completion of the theoretical component delivered within a 3–4 year law degree and the practical legal training program that addresses the professional standards. Assessments within each component may vary depending on where the degree is undertaken and who provides the practical training. Drawing on the progressive outcomes of international reform initiatives, in 2017, the Assuring Professional Competence Committee proposed the development of a *Competence Statement for Australian Legal Practitioners*. As discussed in Chapter 1 in this book, a profession's move to embed a competence assessment undertaken in the field – to normalize such assessment – involves culture change and takes time.

In each profession discussed above, there is the common recognition of the need to assess theoretical understandings, knowledge, and skills as well as performance in real-world contexts before entry to the profession. Evident in statements of professional standards, such as in those in paramedicine and nursing, against which performance is measured, is also a growing awareness of the need to broaden the scope of assessment to include, for example, the assessment of dispositions, communication, and collaboration. ANSAT, referred to above, is illustrative of the move away from one-off or staged demonstrations for assessing competence to workplace appraisal instruments that capture authentic practice in situ and over a sustained period. However, questions remain about the reliability of assessing complex performance assessments that move beyond knowledge and skills to dispositions and values, and performance within variable contexts assessed by professionals within that context. Across professions, how assessment of professional competence is conceptualized, designed, and implemented, to take account of the valued dispositions within a profession, is an emerging field. The scan of competence assessment in a range of professions has brought to light the complex issues of the nature and function of evidence (or the lack thereof) and the complex relationship between evidence and standards in arriving at a judgment of professional readiness.

Notably missing in the scan were references to moderation (social and statistical) as contributing to systems and processes for quality assurance. These include processes for analyzing evidence to monitor how standards are applied at the point of entering the profession, and the reliability of judgments of profession readiness. Broadly speaking, this omission could reflect the traditional reliance on how academic programs of preparation typically relied on evidence from examinations to establish pass rates and in turn certification. It could also reflect the widely reported disconnect between preparation through an academic program and the practical preparation program in the field. A significant challenge in establishing competence assessments is how to bring the evidence generated in both programs into an overall statement of profession readiness.

More fundamentally, the discussion identifies that much remains to be known about how to conceptualize and design complex performance assessments. There are big questions surrounding how to assess demonstrations of thinking critically, analyzing practice, engaging in professional relationships, reflecting on and delivering responsive quality practice, and evaluating outcomes to inform practice. Added to the complexity are the challenges associated with demonstrations of ethical work and effective and appropriate communication processes in context. Beyond these are critical issues associated with how dispositions can be assessed and the parameters that guide such assessment. These could include, for example, intercultural communication, awareness of bias and values as they influence interactions, and diversity variables including gender.

We now turn to consider the relationship between evidence, standards, and profession readiness within teacher education, drawing on more than two decades of work in TPAs.

Performance assessments in initial teacher education: Examples from the United States

Performance assessments for entry into the teaching profession have been a legislated requirement in the state of California since 1998. The PACT[2], developed in 2001 through the collaboration of 12 institutions, was one response to this legislation. PACT was a summative portfolio-based assessment that required candidates for teaching to analyze and reflect on their planning, teaching, and assessing practices. Duckor et al. (2014) used item response models (IRM) to examine the internal structure validity of the PACT instrument and assess model fit, consistency of results with expectations, and internal consistency reliability. They found that the internal consistency of the PACT instrument was high, with a person separation reliability index of 0.92. They did not investigate judgment reliability. However, other studies that focused on inter-rater reliability for PACT performances that were double scored found that the evidence for inter-rater consistency was poor to moderate (Porter & Jelinek, 2011). Research into the use of PACT as a competence assessment has raised questions about the "use of a single high-stakes assessment for licensing beginning teachers" (Santagata & Sandholtz, 2019, p. 480). While Reagan et al. (2016) found evidence that PACT was being used formatively by preservice teachers and by teacher educators for program design, their findings also raised issues relating to the impact of competence assessments for licensure on limiting program design and teaching practices. Recognizing this threat, the authors advocated for multiple measures of performance being used to inform licensure decisions.

The PACT assessment informed the development of the edTPA[3], developed by Stanford University. Currently, the edTPA is implemented in over 40 states and the District of Columbia. It is used for credentialing and program authorization in some states, and formatively in others. The edTPA is described as a subject-specific TPA that evaluates a common set of teaching principles, teaching behaviors, and pedagogical strategies that focus on specific content learning outcomes related to teachers' readiness to take up professional practice (Stanford Center for Assessment, Learning & Equity [SCALE], n.d.). Its purpose has been described as "to improve teaching quality by assessing and evaluating, in a robust and valid way, not just what teachers know about learning and teaching, but how they enact their practice and use evidence thereof to impact student learning outcomes" (Meuwissen &

Choppin, 2015, p. 20). It focuses on performance and requires candidates to compile a portfolio with lesson plans, student work samples, short video clips, and a lengthy commentary of 40–60 pages (Greenblatt & O'Hara, 2015). The core elements of edTPA – planning, teaching, and assessing – are described as focused on student learning by

- Drawing from students' prior knowledge and experiences as instructional assets
- Representing the subject matter in ways that meet diverse students' needs
- Analyzing classroom interactions and students' work
- Using the results of those analyses to inform ongoing practice. (Meuwissen & Choppin, 2015, p. 6)

Among the intended outcomes of edTPA is its use to demonstrate candidates' readiness for the classroom, and to provide "meaningful and consistent data that can be used to improve and update teacher preparation programs and renew program curriculum" (Stanford Center for Assessment, Learning & Equity [SCALE], n.d, n.p). The edTPA is currently managed by Pearson (see Pearson.com), a move endorsed and supported by the American Association of Colleges for Teacher Education (Gitomer et al., 2021b). Pearson oversees the distribution and scoring of the assessment.

There has been a plethora of research conducted into the use of the edTPA including its validity and reliability as a competence assessment to determine preparedness for classroom practice. The edTPA has been found to be useful in identifying necessary program supports for teacher education candidates (Cash et al., 2019; Rao et al., 2021). While preservice teachers were found to have strong general pedagogic knowledge, Rao et al. (2021) identified program gaps in supporting preservice teachers to make connections to discipline-specific knowledge and teaching strategies. However, preparing preservice teachers only for the skills assessed in the edTPA was found, in some cases, to impede the development of broader professional knowledge and skills (Potter, 2021; Rao et al., 2021). Teaching to successful completion of edTPA has been criticized as limiting the focus on preparing preservice teachers who can improve student learning (Donovan & Cannon, 2018; Swars Auslander et al., 2021).

Inconsistencies in the support provided to preservice teachers, even within programs, has also been identified by Cohen et al. (2020) who highlighted three challenges to edTPA implementation:

> (a) the potential of top-down mandates leading to a lack of clarity about goals and variable implementation across programs, (b) the salience of program context for influencing implementation—highlighted by candidates' divergent views of support available while completing edTPA, and (c) the multiple professional identities of teacher educators that color perceptions of and engagement with edTPA's professionalization goals.
>
> *(p. 20)*

The authors point to the risk that the edTPA may lead to standardization of teacher education. They proffer an alternate response to a performance assessment that is nuanced to site differences and the complexity of ITE systems. In a similar manner, Potter (2021) cited "the absence of arts educators from the development of edTPA" as leading to "inconsistencies within the evaluation of fine arts teacher candidates" (p. 101). The risk here is that inconsistencies in judgment can limit the validity of the instrument and the results that it generates.

Others have questioned the edTPA's ability to establish readiness for classroom practice. Parkes and Powell (2015) critiqued the edTPA as insufficient for identifying the professional readiness of arts preservice teachers. The authors identified several limitations which included the separate scoring of planning, teaching, and assessing as "discrete and isolated tasks" (p. 105) which simplified the complexity of actual teaching practice and the "interaction among these elements" (p. 104). The requirement to submit two ten-minute videos of teaching was also seen as limited evidence of the complex ongoing interactions invested in actual daily teaching practice. Greenblatt (2019) compared the objectives of the edTPA with preservice teachers' and teacher educators' experiences of the assessment. Collectively, the authors highlighted issues with how the edTPA appeared to privilege certain pedagogic practices and the limited contextual information that can be included in the assessment impacting how preservice teachers represent their practice. Others have made the point that while the edTPA can promote deeper and reflective thinking, the summative aspect of the assessment restricted preservice teachers' active inquiry into their practice and decision-making about pedagogic choices (Paugh et al., 2018).

Measures to ensure the reliability of the edTPA have also been queried. Several authors have identified that while scoring of the edTPA was initially conducted using two scorers to grade each assessment, in most cases, some authors have reported only one scorer is used (Gitomer et al., 2021b; Greenblatt & O'Hara, 2015). This practice has been criticized in terms of the dependability of the measure (Gitomer et al., 2021b). Gitomer et al. (2021a) have claimed that the edTPA fails to meet "the fundamental principles and norms of educational assessment" (p. 38). Beyond these claims, Gitomer et al. (2021a) have posed the provocative question *Who's assessing the assessment?* in their 2021 article presenting a cautionary tale about the edTPA. Their question can be read as more widely applicable. They claimed that the "edTPA was using procedures and statistics that were, at best, woefully inappropriate and, at worst, fabricated to convey the misleading impression that its scores are more reliable and precise than they truly are. Our analysis showed why those claims were unwarranted" (p. 39). Even more provocative was Gitomer et al.'s (2021a) suggestion that their published concerns with the edTPA "were so serious that they warranted a moratorium on using edTPA scores for high-stakes decisions about teacher licensure" (p. 39).

The edTPA case as presented by Gitomer et al. (2021a, b) is salutary for Australia in at least three ways. First, after an implementation period of over a decade, including the positive impact mentioned previously, there is the emergence of serious concerns stemming from the reported "apparent violation of basic norms of truthfulness" related to analyses and results (Gitomer et al., 2021a, p. 39). Such concerns point to the need for ongoing systematic and public evaluation of a TPA as a high-stakes assessment instrument for licensure, extending to how it is implemented, analyzed, and results reported. At issue here is the need for continuous monitoring to establish fitness-for-purpose. Second, the concerns raised by Gitomer et al. (2021a, b) remind us that assessment is big business. Assessing for professional credentialing purposes has an attractiveness for edu-businesses seeking to become influential in education across all levels (Lingard et al., 2017). An example in Australia was the introduction in 2016 of compulsory literacy and numeracy tests to be undertaken by all teacher education candidates (Literacy and numeracy test for initial teacher education [LANTITE]; Australian Council for Educational Research [ACER], 2021). Students pay a fee to sit each of the two components of LANTITE, with each re-sit incurring additional cost. Not surprisingly, this has generated a thriving coaching business in literacy and numeracy. Third, is the issue of how a TPA can serve both summative and formative purposes, especially when it is a high-stakes assessment. As discussed in Chapter 3, TPAs as high-stake assessments can influence teacher education programs and candidates' teaching practice.

At the time of writing this book, we highlight that the design and implementation of TPAs in Australia have been invested in the profession as they have responsibility for the scoring, analysis, and reporting of results. This is discussed in more detail in Chapters 7 and 8, with online cross-institutional moderation (CIM-Online™)[4] of the Graduate Teacher Performance Assessment (GTPA®; see Chapter 4 for a description of this assessment)[5] being a distinctive feature in TPA implementation internationally.

Performance assessments in Australian teacher education: The relationship between competence and readiness

In responding to this highly charged and rapidly changing teacher education policy context internationally, we needed to understand what is meant by the term 'classroom ready' and the nature and scope of the evidence that could be gathered to demonstrate 'classroom readiness' to take up responsibility for independent classroom teaching. In Australian ITE, interest in the concept of readiness intensified after the publication of the Teacher Education Ministerial Advisory Group (TEMAG) report (Craven et al., 2014). Alexander (2018), in reviewing the four documents associated with the report, identified divergent views and expectations within the professional and wider community of the skills and knowledge required of graduate teachers. She found a lack of clarity regarding the definition of 'readiness' and identified three broad ways the term was being used: (1) Preparedness to take up classroom practice across any possible context a graduate teacher may be assigned; (2) preparedness for teaching practice with the appropriate level of generic teaching knowledge, skills, and practices; and (3) preparedness for the profession as encompassing the knowledge, practices, and dispositions recognizable by those in the profession. By identifying similarities across the three uses of the term readiness, Alexander proposed that it be defined as "a professional outcome and as a consequence … a shared responsibility and an initial milestone in professional learning" (p. 106).

The concept of classroom readiness in the TEMAG report was also critiqued as being deficient in the sense that it set up a university education "as no more than advanced training for employment" and suggesting "that there is no need for further learning or development throughout a career" (Mills & Goos, 2017, p. 644). Mills and Goos (2017) identified the TEMAG report as failing to acknowledge teaching as a profession that continues to learn over the course of a career, disregarding the diversity of teaching contexts preservice teachers will encounter in Australia. Alexander's (2018) definition of readiness as the initial stage of continued professional learning attends to Mills and Goos' identified concerns with the depiction of readiness in the TEMAG report. Alexander (2018) concluded that readiness for the profession encapsulated readiness for the classroom and for teaching. In this way, she moved beyond skills to be mastered, toward a focus on teaching as a learning profession in which expertise develops over time.

This extended definition of readiness aligns with other definitions that have distinguished readiness as requiring additional evidence above and beyond that demonstrated through professional practice, or practicums (Kameniar, et al., 2017; Tatto et al., 2012). For example, Kameniar et al. (2017) described the "key to classroom readiness" as "the nexus between their academic studies and professional practice knowledge" (p. 66). Mayer (2014) identified evidence focused on the impact of teaching on student learning as a feature that should be included in assessments of readiness. She connected professional decision-making and impact on student learning with evidence of such decision-making over an extended period.

In this book, we understand a profession-ready graduate as one who is ready to enter employment and take responsibility for classroom teaching. This includes being a reflexive teacher who continues to develop their theoretical and practical knowledges and who can use these knowledges to interrogate the impact of their practice in efforts to improve their teaching and student learning.

Conclusion

The discussion in this chapter has highlighted that competence remains a nebulous concept with ongoing attempts within professions to articulate a common and agreed set of performance expectations or competencies. Performance assessments that demonstrate competence are acknowledged as challenging to design and assess, involving the combination of knowledges, skills, and dispositions across any given performance. Advances within the professions discussed in this chapter show the growing move to assess the experience of being a professional and the work of that profession. This involves assessing the knowledge and skills of the profession as well as the more difficult to capture critical thinking, reasoning, and decision-making skills, and the relational and ethical skills. The suggestion has been made that explication of reasoning of professional judgments and decision-making is one way to 'see' and judge competence (Blömeke et al., 2015).

Beyond the complexities of assessment design are issues of reliability and validity necessary to ensure fairness for candidates and meet expected standards of measurement. Across the professions discussed in this chapter, including in teacher education, is the need to address reliable scoring, the fidelity of implementation, the integrity of the assessment, and accurate reporting. Elsewhere in this book, we propose that digital infrastructures provide opportunities to collect, store, and analyze large-scale data that could be used to inform issues of reliability and validity (see Chapters 6, 7, and 8). In addition, we propose that through online cross-institutional moderation (CIM-Online™), teacher educators have access to data that can be used in program review and improvements, noting that this will require developing teacher educators' data literacy.

Notes

1 This option of four years academic program and two years internship is currently being phased out (Psychology Board Ahpra, 2020).

2 Information on the PACT can be found at https://scale.stanford.edu/teaching/pact

3 Information on the edTPA can be found at https://scale.stanford.edu/teaching/consortium

4 Acknowledgment: The online model of cross-institutional moderation (CIM-Online™) was conceptualized and developed in the Institute for Learning Sciences and Teacher Education, Australian Catholic University. For a discussion of CIM-Online™, readers are advised to also see Wyatt-Smith & Adie, 2021.

5 Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

References

Alexander, C. (2018). Conceptions of readiness in initial teacher education: Quality, impact, standards and evidence in policy directives. In C. Wyatt-Smith & L. Adie (Eds.), *Innovation and accountability in teacher education: Setting directions for new cultures in teacher education* (pp. 97–113). Springer.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Archer, J., Lynn, N., Coombes, L., Roberts, M., Gale, T., Price, T., & de Bere, S. R. (2016). The impact of large-scale licensing examinations in highly developed countries: a systematic review. *BMC Medical Education*, *16*(1), 212, 1–11. https://doi.org/10.1186/s12909-016-0729-7

Australian Council for Educational Research (ACER). (2021). *Literacy and numeracy test for initial teacher education students*. https://teacheredtest.acer.edu.au/

Australian Nursing Standards Assessment Tool (ANSAT). (2018). *ANSAT resource manual*. https://www.ansat.com.au/home/resources

Blömeke, S., Gustafsson, J. E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift Für Psychologie*, *223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194

Bone, A., & Maharg, P. (2019). *Critical perspectives on the scholarship of assessment and learning in law: Introduction-Legal education assessment in England*. ANU Press. http://doi.org/10.22459/CP01.2019

Bridges, S., Wyatt-Smith, C., & Botelho, M. (2017). Clinical assessment judgements and 'Connoisseurship': Surfacing curriculum–wide standards through transdisciplinary dialogue. In D. Carless, S. M. Bridges, C. K. Y. Chan, & R. Golfcheski (Eds.), *Scaling up assessment for learning in higher education* (pp. 81–98). Springer.

Broadfoot, P. (2010). Signs of change: Assessment past, present and future. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice*. (pp. v–xi). Springer. https://doi.org/10.1007/978-1-4020-9964-9

Cash, A., Putman, S. M., Polly, D., & Byker, E. J. (2019). Candidate and program characteristics associated with EdTPA performance. *Action in Teacher Education*, *41*(3), 229–248. https://doi.org/10.1080/01626620.2019.1600602

Cohen, J., Hutt, E., Berlin, R. L., Mathews, H. M., McGraw, J. P., & Gottlieb, J. (2020). Sense making and professional identity in the implementation of edTPA. *Journal of Teacher Education*, *71*(1), 9–23. https://doi.org/10.1177/0022487118783183

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation*, *13*(5), 401–434. https://doi.org/10.1080/13803610701728311

Craven, G., Beswick, K., Fleming, J., Fletcher, T., Green, M., Jensen, B., Leinonen, E., & Rickards, F. (2014). *Action now: Classroom ready teachers*. Teacher Education Ministerial Advisory Group, TEMAG. Department of Education. Australia. https://docs.education.gov.au/documents/action-now-classroom-ready-teachers-report

Diederich, P. B. (1974). *Measuring growth in English*. National Council of Teachers of English.

Donovan, M. K., & Cannon, S. O. (2018). The university supervisor, edTPA, and the new making of the teacher. *Education Policy Analysis Archives*, *26*(28), 1–26.

Downing, S. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Dreyfus, H. L., & Dreyfus, S. E. (2005). Peripheral vision: Expertise in real world contexts. *Organization Studies 26*(5), 779–792. https://doi.org/10.1177/0170840605053102

Duckor, B., Castellano, K. E., Téllez, K., Wihardini, D., & Wilson, M. (2014). Examining the internal structure evidence for the Performance Assessment for California Teachers: A validation study of the elementary literacy teaching event for Tier I teacher licensure. *Journal of Teacher Education*, *65*(5), 402–420. https://doi.org/10.1177/0022487114542517

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *Jama*, *287*(2), 226–235. https://doi.org/10.1001/jama.287.2.226

Falender, C. A., & Shafranske, E. P. (2007). Competence in competency-based supervision practice: Construct and application. *Professional Psychology, Research and Practice*, *38*(3), 232–240. https://doi.org/10.1037/0735-7028.38.3.232

Feltovich, P., Prietula, M., & Ericsson, K. (2006). Studies of expertise from psychological perspectives. In K. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 41–68). Cambridge University Press. https://doi.org/10.1017/CBO9780511816796.004

Freebody, P., & Wyatt-Smith, C. (2004). The assessment of literacy: Working the zone between system and site validity. *Journal of Educational Enquiry*, *5*(2), 30–49.

Gitomer, D. H., Martínez, J. F., & Battey, D. (2021a). Who's assessing the assessment? The cautionary tale of the edTPA. *Phi Delta Kappan*, *102*(6), 38–43. https://doi.org/10.1177/0031721721998154

Gitomer, D. H., Martínez, J. F., Battey, D., & Hyland, N. E. (2021b). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal*, *58*(1), 3–31. https://doi.org/10.3102/0002831219890608

Gonsalvez, C. J., & Crowe, T. P. (2014). Evaluation of psychology practitioner competence in clinical supervision. *American Journal of Psychotherapy*, *68*(2), 177–193. https://doi.org/10.1176/appi.psychotherapy.2014.68.2.177

Greenblatt, D. (2019). Conflicting perspectives: A comparison of edTPA intended outcomes to actual experiences of teacher candidates and educators in New York City schools. *Journal of Inquiry and Action in Education*, *10*(1), 68–90. https://digitalcommons.buffalostate.edu/jiae/vol10/iss1/3

Greenblatt, D., & O'Hara, K. (2015). Buyer beware: Lessons learned from edTPA implementation in New York State. *Teacher Education Quarterly*, *42*(2), 57–67. ISSN: 0737-5328

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice*. Oxford University Press.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245–270. https://doi.org/10.1080/02671520500193744

Health Practitioner Regulation National Law Act 2009 (Cth). https://www.legislation.qld.gov.au/view/whole/pdf/inforce/current/act-2009-045

Heap, J. (1987). *Effective functioning in daily life: A critique of concepts and surveys of functional literacy* [Conference paper]. National Reading Conference, Arizona.

Hoskins, B., & Deakin Crick, R. (2010). Competences for learning to learn and active citizenship: Different currencies or two sides of the same coin? *European Journal of Education*, *45*(1), 121–137. https://doi.org/10.1111/j.1465-3435.2009.01419.x

Kameniar, B., McLean Davies, L., Kinsman, J., Reid, C., Tyler D., & Acquaro, D. (2017). Clinical praxis exams: Linking academic study with professional practice knowledge. In M. A. Peters, B. Cowie, & I. Menter (Eds.), *A companion to research in teacher education* (pp. 53–67). Springer.

König, J., Blömeke, S., Jentsch, A., Schlesinger, L., née Nehls, C. F., Musekamp, F., & Kaiser, G. (2021). The links between pedagogical competence, instructional quality, and mathematics achievement in the lower secondary classroom. *Educational Studies in Mathematics*, *107*(1), 189–212. https://doi.org/10.1007/s10649-020-10021-0

Kramer, M., Förtsch, C., Stürmer, J., Förtsch, S., Seidel, T., & Neuhaus, B. J. (2020). Measuring biology teachers' professional vision: Development and validation of a video-based assessment tool. *Cogent Education*, 7(1), 1–28. https://doi.org/10.1080/2331186X.2020.1823155

Law Admissions Consultative Committee. (2015). *Practical Legal Training: Competency standards for entry-level lawyers*. https://www.lawcouncil.asn.au/files/web-pdf/LACC%20docs/Competency_Standards_for_Entry-Level_Lawyers_-_1_July_2015.pdf

Lingard, B., Sellar, S., & Lewis, S. (2017). Accountabilities in schools and systems. In G. Noblit (Ed.), *Oxford Research Encyclopedia of Education* (pp. 3–27). Oxford University Press.

Maharg, P., & Webb, J. (2019). Of tails and dogs: Standards, standardisation and innovation in assessment. In A. Bone & P. Maharg (Eds.), *Critical perspectives on the scholarship of assessment and learning in law* (pp. 25–49). Australian National University Press. http://doi.org/10.22459/CP01.2019

Malik, K. (2014). *Human development report 2014: Sustaining human progress: Reducing vulnerabilities and building resilience*. United Nations Development Programme, New York. http://hdr.undp.org/sites/default/files/hdr14-report-en-1.pdf

Maude, P., Livesay, K., Searby, A., & McCauley, K. (2021). Identification of authentic assessment in nursing curricula: An integrative review. *Nurse Education in Practice*, *52*(March 2021), 1–9. https://doi.org/10.1016/j.nepr.2021.103011

Mayer, D. (2014). Forty years of teacher education in Australia: 1974–2014. *Journal of Education for Teaching*, *40*(5), 461–473.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741–749. http://psycnet.apa.org/journals/amp/50/9/741.pdf&uid=1996-10004-001&db=PA

Meuwissen, K. W., & Choppin, J. M. (2015). Preservice teachers' adaptations to tensions associated with the edTPA during its early implementation in New York and Washington States. *Education Policy Analysis Archives*, *23*(103), 1–25. doi:10.14507/epaa.v23.2078

Mills, M., & Goos, M. (2017). The place of research in teacher education? An analysis of the Australian Teacher Education Ministerial Advisory Group Report *Action Now: Classroom Ready Teachers*. In M. A. Peters, B. Cowie, & I. Menter (Eds.), *A companion to research in teacher education* (pp. 637–650). Springer.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5–12. https://doi.org/10.3102/0013189X023002005

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. SAGE Publications.

Nursing and Midwifery Board AHPRA. (2021). *Registered nurse standards for practice*. https://www.nursingmidwiferyboard.gov.au/Codes-Guidelines-Statements/Professional-standards/registered-nurse-standards-for-practice.aspx

Ossenberg, C., Mitchell, M., & Henderson, A. (2020). Adoption of new practice standards in nursing: revalidation of a tool to measure performance using the Australian registered nurse standards for practice. *Collegian*, *27*(4), 352–360.

Paramedic Competency Assessment Consortium. (n.d.). *Candidate handbook: AHPRA paramedic competency assessment*. https://static1.squarespace.com/static/5c4112a8697a98cc463b6ea5/t/5d009fb326b4cb000133e1af/1560321976828/Candidate+Handbook.pdf

Paramedicine Board of Australia, Australian Health Practitioner Regulation Agency (AHPRA). (2021). *Professional capabilities for registered paramedics*. https://www.paramedicineboard.gov.au/professional-standards/professional-capabilities-for-registered-paramedics.aspx

Parkes, K. A., & Powell, S. R. (2015). Is the edTPA the right choice for evaluating teacher readiness? *Arts Education Policy Review*, *116*(2), 103–113. https://doi.org/10.1080/10632913.2014.944964

Paugh, P., Wendell, K. B., Power, C., & Gilbert, M. (2018). 'It's not that easy to solve': edTPA and preservice teacher learning. *Teaching Education*, *29*(2), 147–164.

Popham, W. J. (2017). Disposed to measure dispositions. *Teacher Educator*, *52*(3), 284–289. https://doi.org/10.1080/08878730.2017.1316606

Porter, J. M., & Jelinek, D. (2011). Evaluating inter-rater reliability of a national assessment model for teacher performance. *International Journal of Educational Policies*, *5*(2), 74–87. ISSN: 1307-3842.

Potter, D. (2021). (In)visible power: A critical policy analysis of edTPA. *Arts Education Policy Review*, *122*(2), 101–114. https://doi.org/10.1080/10632913.2020.1744053

Psychology Board Australian Health Practitioner Regulation Agency (AHPRA). (2020). *General registration*. https://www.psychologyboard.gov.au/Registration/General.aspx

Psychology Board of Australia. (2016). *Registration standard for general registration*. https://www.psychologyboard.gov.au/Standards-and-Guidelines/Registration-Standards.aspx

Psychology Board of Australia. (2017). *Registration standard: Provisional registration*. https://www.psychologyboard.gov.au/Standards-and-Guidelines/Registration-Standards.aspx

Psychology Board of Australia. (2019). *Guidelines for the national psychology exam*. https://www.psychologyboard.gov.au/Standards-and-Guidelines/Codes-Guidelines-Policies/Guidelines-for-national-psychology-exam.aspx

Rao, A. B., Olson, J. D., & Koss, M. D. (2021). Teacher candidate perspectives of edTPA readiness: Preparation programs and edTPA supports. *Teaching Education*, *32*(3), 251–268. https://doi.org/10.1080/10476210.2020.1713079

Reagan, E. M., Schram, T., McCurdy, K., Chang, T. H., & Evans, C. M. (2016). Politics of policy: Assessing the implementation, impact, and evolution of the Performance Assessment for California Teachers (PACT) and edTPA. *Education Policy Analysis Archives*, *24*(9), 1–24. https://doi.org/10.14507/epaa.24.2176

Sadler, D. R. (1985). The origins and functions of evaluative criteria. *Educational Theory*, *35*(3), 285–297. https://doi.org/10.1111/j.1741-5446.1985.00285.x

Santagata, R., & Sandholtz, J. H. (2019). Preservice teachers' mathematics teaching competence: Comparing performance on two measures. *Journal of Teacher Education*, *70*(5), 472–484. https://doi.org/10.1177/0022487117753575

Smith, A. C., Framp, A., & Andersen, P. (2020). Assessing competence of undergraduate paramedic student practice: A preliminary evaluation of the Australasian Paramedic Competency Assessment Tool. *Australasian Journal of Paramedicine*, *17*, 1–8. https://doi.org/10.33151/ajp.17.804

Stanford Center for Assessment, Learning and Equity [SCALE]. (n.d.). About edTPA: Outcomes. http://edtpa.aacte.org/about-edtpa#Overview-0

Swars Auslander, S., Meyers, B., Tanguay, C., Smith, S. Z., & Myers, K. D. (2021). High-stakes assessment in an elementary teacher preparation program: A case study of multiple stakeholders. *Teacher Development*, *25*(3), 366–388. https://doi.org/10.1080/13664530.2021.1915371

Takashima, M., Burmeister, E., Ossenberg, C., & Henderson, A. (2019). Assessment of the clinical performance of nursing students in the workplace: Exploring the role of benchmarking using the Australian Nursing Standards Assessment Tool (ANSAT). *Collegian*, *26*(4), 502–506. https://doi.org/10.1016/j.colegn.2019.01.005

Tatto, M., Schwille, J., Senk, S., Ingvarson, L., Rowley, G., Peck, R., … Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)*. International Association for the Evaluation of Educational Achievement (IEA). https://files.eric.ed.gov/fulltext/ED542380.pdf

Wiliam, D. (1994). Reconceptualising validity, dependability and reliability for national curriculum assessment. In D. Hutchison & I. Schagen (Eds.), *How reliable is national curriculum assessment?* (pp. 11–34). National Foundation for Educational Research. https://www.nfer.ac.uk/media/1422/91098.pdf

Wyatt-Smith, C., & Adie, L. (2021). Introducing a new model for online cross-institutional moderation. In C. Wyatt-Smith, L. Adie, & J. Nuttall (Eds.), *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration* (pp. 39–58). Springer.

3

THREE ESSENTIAL FEATURES IN A NEW CONCEPTUALIZATION OF COMPLEX PERFORMANCE ASSESSMENTS

Authenticity, system and site validity, and intelligent accountability

Introduction

A hallmark of a profession is how it takes responsibility for itself: Promoting agency and mutual responsibility among those who claim professional membership. This chapter is concerned with the utility of complex performance assessment for determining entry to teaching. In addressing this topic, we were mindful of the recurring refrain that performance assessments have the potential to standardize practice and in so doing, reduce the agency of teacher educators and teachers. In countering this, our approach was to build a community of professional responsibility and trust.

We drew on three concepts: *Authenticity*, *system and site validity* (Freebody & Wyatt-Smith, 2004), and *intelligent accountability* (O'Neill, 2002, 2013). These formed the foundation of a new competence assessment for teaching in Australia, known as the Graduate Teacher Performance Assessment (GTPA®).[1] All three concepts have existed for some time in the assessment literature. We draw on these to address assessment *fitness-for-purpose*. They provide an opening to see professional readiness, first, through a lens to see that the assessment provides opportunities for demonstrating the knowledge and skills recognized by the profession. This is to say, the performance of the assessment applies the knowledge and skills in real-world (authentic) contexts that include the messiness of interactions in context. Consider, for example, interactions with clients, patients, and in the case of teaching, students, families, and other staff.

Second is the lens to see the interactions between the system and site expectations of practice. The former are typically specified in governance structures and required by professional associations, for example, standards of practice and related professional registration. The latter are the practices and interactions that become normalized in a site over time. These are normally well established before the arrival of a candidate undertaking a professional placement on site, for example, in a dental

DOI: 10.4324/9780429318504-4

clinic, hospital, legal office, or school. Of special interest is the potential for tension between the two sets of expectations: System and site.

Third, with a focus on intelligent accountability, we argue that complex competence assessments should be carried out by those with appropriate evaluative experience and expertise and current knowledge of practice in the field. Through the GTPA Collective of universities, we also seek to develop systems of mutual responsibility and trust as essential to the professionalism of teacher educators.

In the next section, we unpack each of these concepts and show how they relate, each to the other, in a new coherent whole for conceptualizing performance assessments. We also present eight research-informed design principles informed by this conceptualization and apply them in the design of the GTPA.

Authenticity as a design feature of complex performance assessment

Authentic assessment and *competence assessment* are often described in the literature with reference to similar features. Authentic assessments within professional contexts have been described as replicating those practices expected of one working within the field (Gulikers et al., 2004). *Professional competence* has been described "along a continuum from traits (cognitive, affective, motivational) that underlie the perception, interpretation, and decision-making that give rise to observed behavior in a particular real-world situation" (Blömeke et al., 2015, p. 11), as well as "the ability to use professional knowledge and skills to solve problems that arise in practice" (Kane, 1992, p. 164). Thus, professional competence involves being able to (1) "handle the encounters" (Kane, 1992, p. 165) in a professional domain through (2) professional tools (knowledge and skills) combined in various ways to meet client needs. Authentic assessments focus on those knowledges, skills, and dispositions identified with, and accepted and recognized by, the profession. Authentic assessments based on defined professional competences have been described as creating "opportunities for students to integrate learning and working in practice … [resulting] in students' mastery of professional skills needed in their future workplace" (Koh, 2017, p. 2).

Broadly speaking, authentic teaching assessments will capture teaching practice enacted in ways recognizable to the profession. Professional competences generally expected of contemporary teachers include data literacies (Brown et al., 2017) and instructional practices that address 21st-century learning priorities. These include complex decision-making and problem-solving, working individually and in teams (Care et al., 2016), and literacy and numeracy instruction as foundational to students' academic success (Geiger et al., 2015; Wyatt-Smith & Cumming, 2003). The assessment field has known for some time the importance of developing teachers who are data-savvy and who can collect and use data from a range of sources in their instructional decision-making (Black & Wiliam, 1998; Delandshere, 2002; Sadler, 1987). Sadler (1987) identified the assessment dispositions that teachers need to guide their practice, including their willingness to develop their students' evaluative experience and expertise. Black and Wiliam

(1998) identified the use of assessment to inform instruction and to provide effective feedback to students on how to improve their learning.

Delandshere (2002) proposed that assessment be considered as an inquiry into learning. Since this time, assessment scholars have acknowledged "the growing imperative for teachers … to be assessment and data literate" (Cowie & Cooper, 2016, p. 159), and have attempted to describe the diverse range of required knowledges, skills, and dispositions for effective practice (DeLuca et al., 2016; Mandinach & Gummer, 2016). Thus, teachers should have the knowledge and skills to discern and use relevant data to recognize, identify, and analyze students' learning needs and the related barriers to successful learning (Mandinach et al., 2015). They should have the ability to use this knowledge to plan an instructional sequence. Further, as different types of data are collected, they need expertise in inferring meaning from them and adapting their plans to meet students at the point of cognition to guide their learning (Black & Wiliam, 2018; Deakin Crick et al., 2007).

This understanding of teaching and assessment puts learning at the center, with both understood as situated, interactive practice: It acknowledges the 'in-the-moment decision-making' that is responsive to the two-way flow of feedback during teaching (Schoenfeld, 2014). Digital technologies have contributed to rethinking traditional understandings of feedback as transmissive from the teacher to the student, focusing instead on dialogic and active student engagement with feedback. These developments have enabled different ways for teachers to communicate with students and parents, and for more timely feedback than was possible in previous eras (Adie et al., 2018; Van der Kleij & Adie, 2018).

A claim that a TPA is an authentic assessment can be tested against whether it is recognizable by the profession as the work of the profession. Planning, teaching, assessing, and reflecting are recognizable core practices in teaching internationally; they are also core practices in the Australian requirements for a TPA (AITSL, 2017). We propose that the complexity of actual teaching practices requires a TPA that goes beyond the understanding of core practices as discreet components or as linking in a neat linear or cyclical manner. Teaching and learning are complex and messy. Authentic practice is realistically represented as a shuttling or to and fro movement across the practices, as information from a range of sources is continuously collected, synthesized, and evaluated (Wyatt-Smith & Gunn, 2010). In these processes, meaning is inferred from the evidence, related to student learning, and then acted upon.

The GTPA is an authentic assessment of teaching. It extends beyond the content of what is taught to include a focus on instructional decision-making. Such decision-making is to be informed by evidence of learning, situated in a particular school and wider community context. By acting on collected data, preservice teachers interrogate the assumptions they make about students' prior learning and readiness to proceed. This is a recursive practice which requires them to update, incidentally and formally, what they know about classroom students' learning and dispositions and how these impact participation and engagement in content areas. Central to effective practice therefore is a preservice teacher's willingness to enquire into a student's current level of performance, to negotiate realistically attainable goals, and monitor progress with students as they move toward higher performance levels.

While teaching occurs in the moment, a teacher's actions in the classroom are inevitably factoring in learning futures for students. These provide a basis for developing learning goals and teaching strategies. Authentic assessments of teaching have been identified as assisting preservice teachers to link theory and practice, taking theoretical generalizations to descriptions of their manifestation in a particular site (Darling-Hammond & Snyder, 2000). In this way, authentic assessments may bridge the disconnect between educational theory, research and policy, and teaching practice which has been highlighted repeatedly in the literature (Biesta, 2007; Smith, 2007; Vanderlinde & van Braak, 2010). For example, Vanderlinde and van Braak (2010) stated that teachers perceive a lack of applicability of theory to their practice. Smith (2007) concluded that there "seems to be more to teaching than the product of theoretical knowledge and practical skills" (p. 281) and that teaching occurs "in response to specific situations within unique contexts" (p. 282).

Optimally, the design of a TPA promotes the explication of situated practice through theory, not in a simplistic 'causal' sense, but as a justification of teaching decisions that are responsive to multiple and varied contextual demands (Biesta, 2007). By linking theory and practice, authentic assessments capture teaching as an inquiry process (Delandshere, 2002; Wyatt-Smith & Gunn, 2010), recognizable within a broad conceptualization while drawing illustrations from a local site. As an inquiry into practice, authentic assessments differ from other modes of assessment that require candidates to reproduce information out of context. Authentic assessments of teaching are situated in classroom contexts and involve interactions with students and other teachers. They involve preservice teachers constructing knowledge drawing on theory, research, and policy, as well as historic personal experiences and current teaching practice.

Designing a TPA as fit-for-purpose therefore requires an assessment in which teaching is recognized to be an intellectual pursuit, that "bridges between the universal terms of theory and the gritty particularities of situated practice" (Shulman, 1998, p. 519). Attention to both actions and informed decision-making are described as activating reflection "to make contact with the core qualities which are of importance at that particular moment" (Korthagen & Vasalos, 2005, p. 68). In the GTPA, preservice teachers provide critical analysis, justification, and defense of their pedagogic decision-making (see Chapter 4 for a discussion of the design of the GTPA). The focus here is on the combination of 'what' action and the 'why' of teaching: What action/strategy can be usefully taken in this teaching event and related interactions to extend student learning and support engagement? Why is one strategy/practice better to use than another for this child or group of children? Engaging in teaching as inquiry entails well-developed content and pedagogic knowledge, which is evidenced, reflected on, and discussed through a range of theoretical lenses (Darling-Hammond & Snyder, 2000; Hurtado, 2001; Shulman, 1998).

Capturing both the process and product of professional decision-making, including deep engagement with activities, problem-solving capabilities, self-assessment, and self-direction, has been identified as important in authentic assessment design and practice (Koh, 2017). Sadler (2013) emphasized that students need to develop skills in discerning or recognizing quality and use these skills to appraise their performance and choose the necessary actions to make adjustments for improving, even during the process of production. This self-appraisal involves being able to 'see' quality and know strategies to put in place.

Kress (2000) used the metaphor of learning to 'see' as engaging with knowledge and recognizing quality in discipline-specific ways. We extend this metaphor to preservice teachers 'seeing' the effectiveness of their teaching on student learning through mechanisms of self-appraisal and self-assessment. Thus, seeing learning as a teacher involves consideration of the performance of both self and students from the vantage point of changes that have occurred in student engagement over time and whether these connect back to the teaching that has occurred.

Learning to 'see' one's teaching practice involves the ability to critically self-assess within disciplinary specializations and informing pedagogical frameworks, to discern impact on learning and learners. 'Seeing' practice, and especially the decisions that led to changes in practice and student learning, reflects the goal of preparing teachers to be active professionals who view practice as a subject of inquiry, appraise the impact of their teaching on student learning, and respond through reflexive deliberation (Archer, 2007).

System and site validity as a design feature of complex performance assessment

The authentic work of teaching is deeply integrated in the context and needs of communities, and the expectations of the community (Darling-Hammond & Snyder, 2000). Teaching is understood as a cultural and historical practice that is inevitably situated in time and place (Wertsch et al., 1995). Notions of 'good' or 'effective' teaching are susceptible to change over time and vary in different contexts. Here, context extends beyond geography to include the more specific attributes or characteristics of schools. These include, for example, school governance and related philosophy, pedagogical frameworks, style of leadership and related beliefs about behavior management, organizational approaches, and school resourcing.

Such attributes can be conceptualized as occurring within a dynamic and interactive network of influences on a school. These are neither stable nor fixed. They can however affect efforts by school leaders and teachers to achieve coherence between requirements for system validity and site validity (Freebody & Wyatt-Smith, 2004). The term *system validity* refers to the official, endorsed policy requirements for public accountability that are typically expected to remain generalizable across sites. As it applies to TPAs, system validity refers to the characteristics of the assessment and the related scoring rubric that are recognizable to the teaching profession and education policy leaders, a point made earlier in this chapter (Adie & Wyatt-Smith, 2018). In Australia, the knowledge, skills, and capabilities of the Australian Professional Standards for Teachers (APST; AITSL, 2011) and

Accreditation Standards (AITSL, 2015) set the quality assurance and accountability requirements for the system validity of TPAs.

Site validity is concerned with what is valued and comes to be normalized over time in how social practices and interactions occur in local contexts. In the case of teaching, efforts to achieve site validity are evident when school leaders and teachers undertake planning that takes account of the school and surrounding community, as well as the knowledge, skills, attitudes, and dispositions that students bring into the school from the community. Given this, some aspects of site validity may remain unarticulated and are typically learned over time as accepted ways of speaking, acting, and thinking. To achieve site validity, TPAs should provide evidence of enacted practices that take account of salient and specific features of the school-in-community context. These practices, for example, could be those valued in the context for engaging students from diverse cultural and linguistic backgrounds. Readers are advised to see Chapters 4, 6, and 7 for further discussions of validity.

Authentic assessments of teaching must be responsive to, and representative of, both the system and site requirements on teachers and their work. The deliberate bringing together of system validity and site validity in the design of TPAs recognizes the importance of context. On the one hand, the assessment must meet official policy requirements, governance structures, and regulatory requirements. On the other hand, the requirement for authentic assessment means that TPAs are designed to allow for a range of contextual variables responsive to diverse school communities, student needs, subject areas, year levels, school programs, and pedagogic frameworks (Figure 3.1).

Context matters. It is a basic variable in both system validity and site validity and how they co-exist in practice. Both can permeate individual and group decision-making and practices. Teachers well know that expectations can come from within their school and local community. They can also come from agencies external to the community. To the extent that these come together in harmony, then the lived experience is that the expectations for practice, and more specifically, what counts as quality practice, are consistent. To the extent that there is a rub or tension between them, they can be the source of confusion or conflict in understanding what are valued practices at the local level.

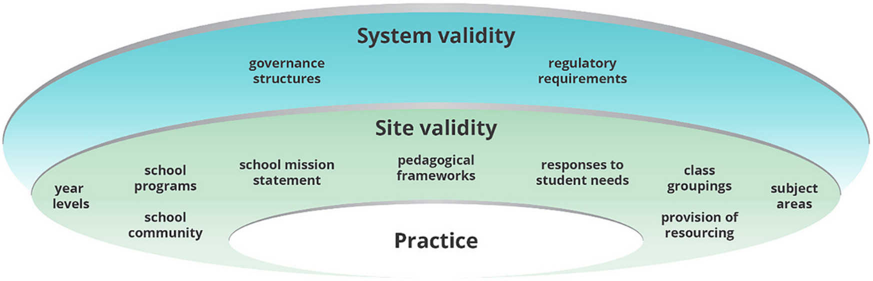


FIGURE 3.1 System and site validity as it permeates practice

A salutary example that illustrates the tensions that can occur between a school's expectations of how teaching will or should occur (site validity) and system expectations of a quality performance on a TPA (system validity) has been provided by Kuranishi and Oyler (2017). The authors identified and described a case of unresolved tensions that were consequential for Kuranishi who initially failed the edTPA reportedly due to "paradigmatic conflicts" (p. 299). They "examine possible explanations for why Adam [Kuranishi] (first author), a New York City public school special educator, failed the edTPA" (p. 299; see Chapter 2 for a discussion of the edTPA). During a year-long teaching residency, Adam achieved "glowing reviews on all program assessments, including 12 clinical observations and firsthand evaluations by his principal and one student" (p. 299). The paradigmatic conflicts could be traced back to the dissonance between (1) the system expectations regarding the use of "the Pearson/SCALE (Stanford Center for Assessment, Learning, and Equity) edTPA expectations or scorer training" and (2) the local site level expectations for teaching in a "heterogeneous Integrated Co-Teaching classroom using Universal Design for Learning and culturally sustaining pedagogy" (p. 299). This example sheds light on the potential regulatory influence of TPA official expectations on preservice teacher performances. It also highlights challenges for achieving dependable and fair scoring (see Chapter 5). Kuranishi and Oyler's (2017) account of unresolved system and site tensions highlights the potential for TPAs to have a standardizing influence on teacher education, preservice teacher practice, and credentialing when system validity is prioritized over site validity.

Throughout the process of developing the GTPA (see Chapter 4), the research team was mindful of repeated assertions that TPAs would standardize teacher education and would be an exercise in compliance. In our experience, standardization has not occurred. This reflects the starting point that teacher education involves a wide range of higher education and school settings, and further that responsiveness to student learning in context needed to be integral in both the design and implementation of the GTPA. A distinctive feature of the GTPA is its purposeful inclusion of system and site as complementary assessment purposes – of measurement and improvement. Chapter 8 describes the use of GTPA scoring data, cumulated cohort data, and program demographic data to inquire into program effectiveness.

Intelligent accountability as a design feature of complex performance assessment

In reflecting on the risks for standardization in the case of the edTPA, Reagan et al. (2016) discussed how responsibility for scoring the edTPA is held by Pearson Education as an external private corporation; Pearson determines "who scores the edTPA, how scorers are trained, how scores are presented, and coordinates where data are housed" (p. 17). The authors highlighted that while scoring decisions "are made in collaboration with the profession (i.e., Stanford Center for Assessment, Learning and Equity [SCALE] and the American Association of Colleges for

Teacher Education [AACTE])", there is the concern that "the inclusion of another external partner in the assessment process creates opportunities for multiple agendas, strategies, and maneuverings outside of the profession" (p. 17). They found evidence of a "potential de-professionalization in state policy contexts" (p. 17). In particular, this occurred in places where teachers and teacher educators had minimal input into the decisions associated with edTPA implementation. These observations point to the critical issue of the roles of the profession and other stakeholders, including edu-business, in the conceptualization, design, and implementation of TPAs.

Raising a similar concern in their discussion of changing modes of governance in Australian teacher education, Savage and Lingard (2018) highlighted the concerns of standardization, de-professionalization, and commercialization. They described the introduction of the APST into the teaching profession as distancing those in the profession from decision-making processes. They went on to assert that, with this move, power has been "increasingly concentrated in the hands of the federal government and AITSL" (p. 78), enabling "the progressive nationalisation and standardisation of policies" (p. 65). This intensified with the policy move that mandated TPAs as a requirement for graduation. In this context, TPAs became a potent mechanism for quality assurance in reforming teacher education. To counterbalance the press for standardization in ITE, the GTPA Collective took up the goal of maintaining responsibility for the profession within the profession. Professional judgment, collaboration, and partnerships were therefore repositioned at the heart of the reform. This stance on professionalizing teacher education aligns with those who have advocated for new forms of accountability (e.g., Cochran-Smith et al., 2017; O'Neill, 2013; Sahlberg, 2010).

Cochran-Smith et al. (2017) used the term democratic accountability to encompass notions of teacher educators' professional responsibility for students' learning, principles of strong equity, and strong collaboration with multiple stakeholders. O'Neill (2013) presented a similar position using the term intelligent accountability in which assessment is carried out "by people who are sufficiently informed to judge the performance they assess, sufficiently independent to do so objectively, and able and permitted to report intelligibly to the various audiences to whom an account is to be given" (p. 15). Sahlberg (2010) identified features of intelligent accountability to include systems of mutual responsibility built on cultures of trust.

Intelligent accountability acknowledges the professional responsibility teachers have for student learning and their personal responsibility for self and community within the profession. When TPAs are viewed through a lens of intelligent accountability, professional responsibility is foregrounded. By retaining responsibility within Australian universities for scoring TPAs and processes of cross-institutional moderation online (CIM-Online™; see Chapter 7),[2] intelligent accountability is enacted within systems of mutual responsibility.

In our work, intelligent accountability is understood to require knowledge of the field (e.g., discipline or content knowledge), knowledge of the practice (e.g., pedagogic knowledge), and evaluative experience and expertise. It is based on the premise that to give an account of practice, you have to know what you are being asked to give an account about and the expectations you have for giving this account. This stance is informed by the notion of professional responsibility (Cochran-Smith et al., 2017), where responsibility and agency are vested in the teaching profession. Intelligent accountability involves giving an account through reflecting on and using the data, bringing to bear local contextual knowledge and experience in the ITE programs that are the focus of the account, to interrogate the data and infer meaning from it. Intelligent accountability recognizes how social–relational cultures shape teaching and assessment in higher education, including teacher education. These cultures develop over time and can have profound effects on opportunities to develop trust, confidence, and mutual responsibility for decision-making, including assessing and scoring student work samples.

Here, we turn to the two main purposes of assessment: (1) The summative purpose where evidence is used for reporting achievement assessed against an agreed established standard at a terminal or juncture point (end of a term or year) and (2) the formative purpose where evidence is used to diagnose learning needs, inform teaching, and progress learning. Traditionally, these purposes have existed as a dualism though in recent decades there has been a concerted attempt to reframe these purposes through the body of writing by assessment scholars (e.g., Black & Wiliam, 2018; Stobart, 2008). Added to this are the concepts of feedback and feedforward, especially as these are directly tied to understandings about the nature and function of human judgment, standards, and evaluative experience and expertise (Sadler, 1989).

The stance taken in the GTPA is that the data collected for summative purposes (establishing graduate competence and program impact) can be ascribed a formative purpose and used for curriculum review and program renewal. In these ways, the intent is to achieve intelligent accountability through self-regulation by the profession. This includes, for example, in scoring and CIM-Online™ involving the application of a stated standard that has been accepted by the profession (see Chapter 7). It can also include applying judgment and expertise to discern opportunities for improving teacher preparation programs in local contexts (see Chapter 8). Through intelligent accountability, the data generated from preservice teacher responses to the requirements of a TPA can be used as a catalyst for re-seeing quality with an evidence-informed gaze that has not been available previously to teacher educators.

The next section illustrates how a conceptualization of a TPA as a competence assessment, responsive to the features of authenticity, system and site validity, and intelligent accountability, can inform key design principles of a TPA. In the following section, we draw on the example of the GTPA as an endorsed TPA currently being implemented across several Australian universities, to illustrate the connection from conceptualization to a set of design principles.

Designing the Graduate Teacher Performance Assessment (GTPA) for authenticity, system and site validity, and intelligent accountability

The GTPA has been designed with a sustained focus on preservice teachers' instructional decision-making through the collection and analysis of the evidence of student learning. It is designed as an authentic assessment of teaching as interconnected practices. Specifically, the GTPA is designed to capture the following:

1. Planning of teaching practices through the constructive use of student data to inform teaching and learning, including the explicit teaching of literacy and numeracy in the curriculum.
2. Teaching decisions regarding the best-suited pedagogies to offer differentiated learning opportunities within a particular teaching context.
3. Assessment design, feedback, and moderation processes in relation to the quality of students' work.
4. Reflections on teaching and learning, drawing on research and theory to support decisions.
5. Critical appraisal of teaching and learning, informed by the appropriate use of student data.

The Australian designed GTPA has several design features in common with the U.S. designed edTPA, as described by Sato (2014). Significantly, both (1) place student learning at the center of the design; (2) focus on the core pedagogic practices of planning, teaching, and assessing; (3) are based on the ongoing collection and analysis of data; (4) aim to capture pedagogic decision-making; (5) consider the broad impact teaching practice may have on student learning; (6) emphasize the requirement for evidence of a variety of instructional and assessment practices to address diverse student needs; (7) require preservice teachers to identify how their theoretical knowledge has informed their classroom decision-making; and (8) present whole class teaching over a sustained period of time, including detailed information from representative focus students.

Particular to the Australian GTPA are (1) the direct link to the established national standards for teachers at the graduate level (APST; AITSL, 2011), that is, the GTPA does not define skilled performance – rather it responds to the established professional standards and goes beyond this to establish the boundary between meets and does not meet the required standard of performance; (2) the requirement for preservice teachers to appraise the impact of their teaching on student learning, thus providing evidence of their capacity to critically reflect and inquire into their planning and teaching practices; (3) the provision of one set of guidelines that accommodate different subject specializations as well as school and class contexts; (4) the requirement that the responsibility for judging the assessment must stay within the profession, that is, teacher educators grade the GTPA and so are accountable for monitoring the application of the standard; and (5) the related engagement of teacher educators in moderation of GTPAs ensuring the development of a shared understanding of the standard. The moderation process ensures that feedback regarding the quality of responses is used to inform ongoing program development and reform (see Chapter 8).

The GTPA has been designed to show the connectedness of planning, teaching, and assessing through critical reflection and appraisal processes. It is designed as an authentic assessment to capture teaching as critical inquiry (Delandshere, 2002; Wyatt-Smith & Gunn, 2010). It is a competence assessment in that it assesses research-informed and recognized practices of quality teaching while acknowledging the situatedness of practice in context, responding to system and site validity. In addition, it ascribes to a notion of intelligent accountability in which teacher educators and preservice teachers use data for formative and summative purposes to maintain responsibility for teaching and assessing their students' learning. For example, preservice teachers, in completing the GTPA, show how they have used data to necessitate a change to their intended teaching plan as well as to appraise the impact of their teaching; teacher educators use data from the range of completed GTPAs to identify strengths and weaknesses in their programs and respond to accreditation requirements. Improving practice through the collation and analysis of student learning data and TPA scoring data is integral to the design of the GTPA. For teacher educators, it provides the opportunity for the first time to build an evidentiary base showing the quality of ITE locally, and potentially, nationally.

Design principles of the GTPA

Eight overarching design principles of the GTPA were distilled from the underpinning conceptualization of what it means to be a competence assessment of teaching that addresses features of authenticity, system and site validity, and intelligent accountability. Each principle and its practical implication for the GTPA is briefly described below.

Principle 1: Teaching practice is informed by theory, research and policy, and enacted through an inquiry approach to teaching and assessing. The GTPA has been designed as a presentation of practice, drawing on a final year professional experience while focusing on the justification of pedagogic decision-making. Such an approach involves preservice teachers drawing on educational theory, research, and systemic policy, supported through their curated body of site-specific evidence, to justify teaching decisions in an authentic teaching context. Assessors should be able to see and understand a preservice teacher's reasoning of their decisions and actions within the particular context of their professional experience.

Principle 2: Practices of planning, teaching, assessing, and reflexivity are understood as integrally connected to how teachers work in responding to students' learning needs and interests. Preservice teachers demonstrate the iterative relationship between planning, teaching, and assessing through the continual monitoring of student learning and the in-the-moment decisions of authentic teaching practice. In completing the assessment, preservice teachers work to and fro across these practices, showing how their ongoing collection of data informed their next-step teaching decisions and modifications to their original plans. In addition, preservice teachers provide evidence of the alignment of curriculum, pedagogy, and assessment. This includes how achievement standards of the Australian (or other informing) Curriculum (Australian Curriculum, Assessment and Reporting Authority [ACARA], n.d.) are applied in their planning, differentiated teaching practice, feedback on student work, and moderation discussions when confirming judgment decisions and ensuring the dependability of their judgments (Harlen, 2005).

Principle 3: The focus of teaching practice is on students' learning at the 'center'. The GTPA requires preservice teachers to analyze whole-class learning, providing a clear outline of targeted planning to accommodate diversity within the class. There is an additional in-depth analysis of the learning of three focus students who collectively represent the full range of achievement in the class. Preservice teachers use the collected data to identify student learning needs and determine how to progress learning. Their justification for appropriate levels of challenge is made according to curriculum requirements while responding to the range of identified student learning needs.

Principle 4: Teaching is understood as evidence-informed practice. Preservice teachers are required to integrate multiple sources of data to continually inform their planning and teaching. Data may include records and observations of classroom talk and patterns of interaction; records of main points learned from consultations with individual students, teachers, parents, and/or paraprofessionals; detailed analysis of student work samples; and formative, summative, and standardized test data. To develop authentic, contextualized learning goals related to the needs of the whole class and individual students, the collected data are not restricted to a wholly pre-specified set of requirements.

Principle 5: Professional practice involves continuous monitoring of the quality of one's teaching to discern its impact on learning and learners. In completing the GTPA, preservice teachers are required to critically reflect on their teaching by analyzing and explaining their teaching decisions and appraising the impact of their teaching on student learning. This is characteristic of the profession (Shulman, 1998) and integral within a system of intelligent accountability (O'Neill, 2002, 2013). In this process, preservice teachers discuss how they monitor student learning progress, and their pedagogical responses, returning to theory to support their discussion. This includes an analysis of the effectiveness of the resources and teaching strategies they employed, and the changes or lack of change in student results.

Principle 6: Teaching is understood as cultural–historical practice that is contextually situated. In completing the GTPA, preservice teachers first provide contextual information related to the site of their practicum. Justification of pedagogical decisions is based on discipline-specific knowledge and practices, site-specific elements such as the school philosophy and pedagogical framework, and the learning needs and goals for the three focus students and the whole class. Responses show the integration of responsive pedagogical practice and systemic accountability to progress students' curriculum knowledge and skills. Such a positioning opens opportunities to also acknowledge alternate actions that could be taken in different contexts or circumstances.

Principle 7: Professional accountability for a performance assessment should be the responsibility of the university. The collective recognition of professional accountability is evident in practices that situate ownership of the required standard within the university. This includes a move to national online moderation, for the self-monitoring of the profession by the profession. This process has provided a set of analytics that have been used to inform ITE curriculum review and program planning and ensured that capstone knowledge and skills are being sequentially developed across the ITE programs. It is further evident in new partnerships in ITE that have developed organically through teacher educators' work with the GTPA.

Principle 8: Standards–evidence loop assures dual purposes of determining the quality of graduate performance and program design. A TPA is fit-for-purpose if it connects standards and evidence in ways that serve the dual purposes of quality assuring graduate 'profession readiness' (Ingvarson et al., 2014) and contributing to system processes for assuring program quality. Where these purposes are achieved, the introduction of a TPA can serve the best interests of the profession by informing the necessary work of program review and curriculum renewal, and building an evidence base to show the quality of ITE.

Conclusion

In this chapter, we have explored the necessity for TPAs to be an authentic representation of teaching, incorporating evidence-based practices which are recognizable to the profession. Being an authentic assessment includes having both system and site validity, responsive to systemic requirements as well as idiosyncratic site requirements. We also argued that intelligent accountability, in which teacher educators are responsible for implementing, scoring, and moderating TPAs, is a necessary feature of TPAs if they are to move beyond a summative purpose to being used formatively by the profession for program review and renewal. The conceptualization of authenticity, system and site validity, and intelligent accountability are operationalized in the eight design principles of the GTPA. Chapters 7 and 8 describe the processes of scoring, CIM-Online™, and reporting that generate the evidence to build a culture of trust and to demonstrate responsibility in the preparation of the next generation of teachers through the GTPA.

Notes

1 Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

2 Acknowledgment: The online model of cross-institutional moderation (CIM-Online™) was conceptualized and developed in the Institute for Learning Sciences and Teacher Education, Australian Catholic University. For a discussion of CIM-Online™, readers are advised to also see Wyatt-Smith and Adie (2021).

References

Adie, L., Van der Kleij, F., & Cumming, J. (2018). The development and application of coding frameworks to explore dialogic feedback interactions and self-regulated learning. *British Educational Research Journal*, *44*(4), 704–723. https://doi.org/10.1002/berj.3463

Adie, L., & Wyatt-Smith, C. (2018). Research-informed conceptualization and design principles of teacher performance assessments: Wrestling with system and site validity. In C. Wyatt-Smith & L. Adie (Eds.), *Innovation and accountability in teacher education: Setting directions for new cultures in teacher education* (pp. 115–132). Springer.

Archer, M. S. (2007). *Making our way through the world: Human reflexivity and social mobility*. Cambridge University Press.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (n.d.). *Australian curriculum*. https://www.australiancurriculum.edu.au/

Australian Institute forTeaching and School Leadership (AITSL). (2011). *Australian Professional Standards for Teachers*. https://www.aitsl.edu.au/teach/standards

Australian Institute for Teaching and School Leadership (AITSL). (2015). *Accreditation of initial teacher education programs in Australia: Standards and procedures*. https://www.aitsl.edu.au/docs/default-source/national-policy-framework/accreditation-of-initial-teacher-education-programs-in-australia.pdf?sfvrsn=e87cff3c_28

Australian Institute for Teaching and School Leadership (AITSL). (2017). *Teaching performance assessment*. https://www.aitsl.edu.au/deliver-ite-programs/teaching-performance-assessment

Biesta, G. (2007). Bridging the gap between educational research and educational practice: The need for critical distance. *Educational Research and Evaluation*, *13*(3), 295–301. https://doi.org/10.1080/13664530.2021.1915371

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Blömeke, S., Gustafsson, J. E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift Für Psychologie*, *223*(1), 3–13.

Brown, C., Schildkamp, K., & Hubers, M. D. (2017). Combining the best of two worlds: A conceptual proposal for evidence-informed school improvement. *Educational Research*, *59*(2), 154–172. https://doi.org/10.1080/00131881.2017.1304327

Care, E., Scoular, C., & Griffin, P. (2016). Assessment of collaborative problem solving in education environments. *Applied Measurement in Education*, *29*(4), 250–264. https://doi.org/10.1080/08957347.2016.1209204

Cochran-Smith, M., Baker, M., Burton, S., Chang, W.-C., Carney, M., Fernández, M., Keefe, E., Miller, A., & Sánchez, J. (2017). The accountability era in US teacher education: Looking back, looking forward. *European Journal of Teacher Education*, *40*(5), 572–588. https://doi.org/10.1080/02619768.2017.1385061

Cowie, B., & Cooper, B. (2016). Exploring the challenge of developing student teacher data literacy. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 147–163. https://doi.org/10.1080/0969594X.2016.1225668

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education*, *16*(5–6), 523–545. https://doi.org/10.1016/S0742-051X(00)00015-9

Deakin Crick, R., McCombs, B., Haddon, A., Broadfoot, P., & Tew, M. (2007). The ecology of learning: Factors contributing to learner centred classroom cultures. *Research Papers in Education*, *22*(3), 267–307. https://doi.org/10.1080/02671520701497555

Delandshere, G. (2002). Assessment as inquiry. *Teachers College Record*, *104*(7), 1461–1484.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251–272. https://doi.org/10.1007/s11092-015-9233-6

Freebody, P., & Wyatt-Smith, C. (2004). The assessment of literacy: Working the zone between system and site validity. *Journal of Educational Enquiry*, *5*(2), 30–49.

Geiger, V., Forgasz, H., Goos, M. (2015). A critical orientation to numeracy across the curriculum. *ZDM Mathematics Education*, *47*(4), 611–624. https://doi.org/10.1007/s11858-014-0648-1

Gulikers, J. T. M., Bostiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, *52*(3), 67–86.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245–270. https://doi.org/10.1080/02671520500193744

Hurtado, S. (2001). Linking diversity and educational purpose: How diversity affects the classroom environment and student development. In G. Orfield (Ed.), *Diversity challenged: Evidence on the impact of affirmative action* (pp. 187–203). Harvard Education Publishing Group.

Ingvarson, L., Reid, K., Buckley, S., Kleinhenz, E., Masters, G., & Rowley, G. (2014). *Best practice teacher education programs and Australia's own programs*. Australian Council for Educational Research. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1014&context=teacher_education

Kane, M. T. (1992). The assessment of professional competence. *Evaluation and the Health Professions*, *15*(2), 163–182. https://doi.org/10.1177/016327879201500203

Koh, K. (2017). Authentic assessment. In *Oxford Research Encyclopedia of Education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.22

Korthagen, F., & Vasalos, A. (2005). Levels in reflection: Core reflection as a means to enhance professional growth. *Teachers and Teaching: Theory and Practice*, *11*(1), 47–71. https://doi.org/10.1080/1354060042000337093

Kress, G. (2000). "You've just got to learn how to see": Curriculum subjects, young people and schooled engagement with the world. *Linguistics and Education*, *11*(4), 401–415.

Kuranishi, A., & Oyler, C. (2017). I failed the edTPA. *Teacher Education and Special Education*, *40*(4), 299–313. https://doi.org/10.1177/0888406417730111

Mandinach, E. B., & Gummer, E. S. (2016). Every teacher should succeed with data literacy. *Phi Delta Kappan*, *97*(8), 43–46. https://doi.org/10.1177/0031721716647018

Mandinach, E. B., Parton, B. M., Gummer, E. S., & Anderson, R. (2015). Ethical and appropriate data use requires data literacy. *Phi Delta Kappan*, *96*(5), 25–28.

O'Neill, O. (2002). *A question of trust*. The BBC Reith Lectures. Cambridge University Press.

O'Neill, O. (2013). Intelligent accountability in education. *Oxford Review of Education*, *39*(1), 4–16. https://doi.org/10.1080/03054985.2013.764761

Reagan, E. M., Schram, T., McCurdy, K., Chang, T.-H., & Evans, C. M. (2016). Politics of policy: Assessing the implementation, impact, and evolution of the Performance Assessment for California Teachers (PACT) and edTPA. *Education Policy Analysis Archives*, *24*(9), 1–27.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*(2), 191–209. https://doi.org/10.1080/0305498870130207

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sadler, D. R. (2013). Opening up feedback: Teaching learners to see. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students* (pp. 54–63). Routledge.

Sahlberg, P. (2010). Rethinking accountability in a knowledge society. *Journal of Educational Change*, *11*(1), 45–61. https://doi.org/10.1007/s10833-008-9098-2

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of Teacher Education*, *65*(5), 421–434. https://doi.org/10.1177/0022487114542518

Savage, G. C., & Lingard, B. (2018). Changing modes of governance in Australian teacher education policy. In N. Hobble & B. L. Bales (Eds.), *Navigating the common good in teacher education policy: Critical and international perspectives* (pp. 64–80). Routledge.

Schoenfeld, A. (2014). What makes for powerful classrooms, and how can we support teachers in creating them? A story of research and practice, productive intertwined. *Educational Researcher*, *43*(8), 404–412. https://doi.org/10.3102/0013189X14554450

Shulman, L. (1998). Theory, practice, and the education of professionals. *The Elementary School Journal*, *98*(5), 511–526. https://doi.org/10.1086/461912

Smith, K. (2007). Empowering school- and university-based teacher educators as assessors: A school – university cooperation. *Educational Research and Evaluation*, *13*(3), 279–293. https://doi.org/10.1080/13803610701632109

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. Routledge.

Van der Kleij, F. M., & Adie, L. E. (2018). Formative assessment and feedback using information technology. In J. Voogt, G. Knezek, R. Christensen, & K. Lai (Eds.), *Second handbook of information technology in primary and secondary education* (pp. 601–615). Springer.

Vanderlinde, R., & van Braak, J. (2010). The gap between educational research and practice: Views of teachers, school leaders, intermediaries and researchers. *British Educational Research Journal*, *36*(2), 299–316. https://doi.org/10.1080/01411920902919257

Wertsch, J. V., del Rio, P., & Alvarez, A. (1995). Sociocultural studies: History, action, and mediation. In J. V. Wertsch, P. del Rio, & A. Alvarez (Eds.), *Sociocultural studies of mind* (pp. 1–31). Cambridge University Press.

Wyatt-Smith, C. M., & Cumming, J. J. (2003). Curriculum literacies: Expanding domains of assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(1), 47–59. https://doi.org/10.1080/09695940301690

Wyatt-Smith, C. M., & Gunn, S. (2010). Towards theorising assessment as critical inquiry. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 83–102). Springer.

4

HOW DO WE KNOW THAT TEACHER GRADUATES ARE READY FOR PROFESSIONAL PRACTICE?

Designing an assessment for evidence of readiness

The move to teaching performance assessments

International moves toward the introduction of teaching performance assessments (TPAs) represent an effort to enhance the status of the profession, usually through state legislative requirements (Pecheone & Chung, 2006). This move has opened a multitude of questions relating to quality. These include: Where should the standard 'prepared to teach' be set?; How is the validity of an instrument to capture 'quality' demonstrated?; What conditions are necessary for demonstrating high reliability in scoring performances?

Duckor et al. (2014) discussed the meaning of validity for teacher licensure and asked a critical question: "What are the grounds for believing that a particular licensure and certification score has meaning for the profession?" (p. 402). To this we add, what meaning (and value) will parents/carers and the wider community attach to the scores? The authors argued that "teacher educators are obligated to know about the psychometric qualities of the assessments they demand for licensure" (p. 402) and "have a professional responsibility to engage with and monitor the validity evidence for any large-scale testing and examination system" (p. 403).

The argument presented by Duckor et al. (2014) has relevance in discussions of performance assessments and their relationship to professionalization in teacher education. While a continuing question in teaching relates to defining teaching and teacher quality (see discussion in Chapter 1), we are scoping this question to focus specifically on the expectations of graduate teacher competence on entry to the profession. At issue is: *How do we know that teacher graduates are adequately prepared and ready for professional practice?*

That this question continues to be asked reflects that an ITE evidence base (large-scale and longitudinal) to address complex issues of quality is in its infancy. This has led to a situation where universities have been operating "partially blindfolded" in

DOI: 10.4324/9780429318504-5

implementing their initial teacher education (ITE) programs (Cochran-Smith et al., 2013, p. 13). They have lacked suitable data to inform them how effective their programs are in preparing preservice teachers for classroom practice. Such an evidence base is distinguishable from information about student satisfaction, retention, and progression. Rather, the evidence we refer to has not been collected previously, and scant attention has been given to how we measure preparedness for practice at the point of graduation. Further, the issue of evidence had not been chartered in systematic ways prior to the arrival of the Teacher Education Ministerial Advisory Group (TEMAG) review (Craven et al., 2014).

The Australian context for assessing professional competence and readiness for teaching

While TPAs have existed in the United States for almost two decades (Sato, 2014), the introduction of TPAs in Australia represented a radical move, the repercussions of which are only now becoming apparent some six years later. In Australia, a recognized catalyst for ITE reform was the Federal Government's endorsement of the recommendations of the TEMAG report, with Recommendation 6 calling for:

> higher education providers to demonstrate that their programs have evidence-based pedagogical approaches, effective integration of professional experience, rigorous and iterative assessment of pre-service teachers throughout their education, and *final assessments* that ensure pre-service teachers are *classroom ready*. Higher education providers provide a set of measures that assess the effectiveness of their programs in achieving successful graduate outcomes.
>
> *(Craven et al., 2014, p. xiv [italics added])*

This recommendation heralded the turn to a concerted focus on evidence to show program effectiveness. Other reports in Australia and internationally had identified some of the additional areas (see Chapter 1), namely evidence-based pedagogical approaches, improved integration of the academic program and the school-based practical program, and improved assessment processes. The new probe into program effectiveness came in the form of final assessments to ensure classroom ready graduates. The potency of this suite to stimulate reform is only beginning to be realized some six years after the report was accepted.

The report presented as 'key findings of fact' that "national standards are weakly applied" (p. xi) to ITE programs, and that "initial teacher education providers are not rigorously or consistently assessing the classroom readiness of their pre-service teachers against the Professional Standards" (Craven et al., 2014, p. xi). The call to action was strongly worded, stating that "Providers do not follow a transparent or consistent framework for assessment of classroom readiness and are not held accountable for the quality of their assessment. The Advisory Group believes that this lack of accountability allows providers to graduate pre-service teachers who do not meet the Graduate level of the Professional Standards" (Craven et al., 2014,

p. 33). To strengthen the accountability of providers and assure graduate preparedness, the TEMAG report called for:

1. "ongoing monitoring and examination of the impact of programs on teacher capability and effectiveness essential to continuous improvement and quality assurance. Programs that do not produce effective teachers should not continue to operate. There is significant evidence of system failure in this context" (p. xii)
2. "rigorous assessment of classroom readiness [that] needs to involve providers and schools working in partnership throughout initial teacher education programs. This includes determining the pre-service teacher's ability to effectively integrate theory and teaching practice and assisting them to collect supporting evidence" (p. 33)
3. "a sufficient and up-to-date benchmark of the expectations of graduates entering the profession" (p. 33).

Over the period of the TEMAG review and to the present, assessing readiness for the profession has been, and continues to be, hotly contested (see Chapter 2 for a discussion of readiness; Mills & Goos, 2017). It remains currently unresolved in Australian education ITE policy and practice. While it is broadly recognized that assessment of preservice teacher performance is a shared responsibility between school and ITE providers, the TEMAG report highlighted tensions in how this is enacted, especially as it concerns assessment outcomes. At issue is how classroom teachers, school leaders, and teacher educators contribute to assuring "readiness for the profession" (Craven et al., 2014, p. 31). Such tensions reflected "limited integration of assessment between on-campus and in-school learning" (Craven et al., 2014, p. 31) and fueled a call for benchmarking graduate outcomes across providers and schools.

Surfacing among the myriad of concerns with teacher preparation expressed in the TEMAG report, the word 'benchmark' came to the fore and was profiled as a key means to build quality assurance processes and, in turn, public confidence in program quality and teacher education graduates. As shown in Table 4.1, the word 'benchmark' appeared coupled with standards and used in multiple ways that ranged from (1) benchmarking using the Australian Professional Standards for Teachers (APST; Australian Institute for Teaching and School Leadership [AITSL], 2011) and accreditation standards (AITSL, 2015), (2) a benchmark to establish graduate preparedness and accreditation processes, and beyond this, (3) to benchmark ITE programs against recognized international best practices (Table 4.1). Standards as quality assurance tools and input measures have been available for some time as national approaches to registration and career pathway. Similarly, accreditation standards have been used as national approaches to program quality. However, a common standard or benchmark of graduate preparedness had not been established. Tensions between schools and providers in assessing professional performance are ongoing. Also unresolved is the place of international benchmarking and how this may be conducted to inform reviews of the quality of teacher education in Australia.

TABLE 4.1 Profiling the uses of the term 'benchmark' and associated purposes in the TEMAG review (Craven et al., 2014)

Quality assurance tool	*TEMAG references to benchmarking purposes*	*Available for use*
1. Standards as inputs: a. Professional standards. b. Accreditation standards.	The Professional Standards provide the benchmarks used in national approaches to accreditation of teacher education programs, registration of teachers for employment and formal recognition of the higher level skills of Highly Accomplished and Lead Teachers (p. 3) Quality assurance for initial teacher education programs in Australia from 2013 is through the Accreditation Standards (see Appendix E), which set the benchmark for the quality of the programs offered by 48 higher education providers delivering initial teacher education across the country (p. 7)	Yes
2. Standards as outputs: Quality indicator of graduate performance for professional practice.	… to provide a sufficient and up-to-date benchmark of the expectations of graduates entering the profession… the importance of regularly reviewing and updating the Graduate level of the Professional Standards to ensure the currency of the skills, knowledge and capabilities required for beginning teachers (p. 33) Consistent and transparent graduate assessment against an agreed benchmark is a key feature of profession entry requirements both internationally and in comparable professions in Australia (pp. xix, 32) … the community must have confidence that the benchmarks and processes that assure the quality of programs will drive improvement and will be rigorously applied (p. 1) … a need for moderation of benchmarks for assessing pre-service teachers across providers and schools (p. 31)	No

(*Continued*)

TABLE 4.1 (Continued)

Quality assurance tool	*TEMAG references to benchmarking purposes*	*Available for use*
3. International benchmarking: a. To identify best practices in professional experience in schools b. To address the paucity of information about the performance of teacher education programs in Australia. c. To identify programs which positively impact student outcomes was problematic.	International benchmarking of best practice has identified that staff leading and supervising professional experience in schools should be exemplary teachers who have undertaken focused training for their roles (p. 6) … Australian programs against high-performing international programs known to impact positively on student outcomes was problematic (p. 41)	Developing

While Australia has had professional and program standards for benchmarking purposes for more than a decade, these have not stopped the revolving door of ITE reviews, nor is there evidence that the reviews have built public confidence in the quality of graduates (see Chapter 1). This could reflect several factors. Like many countries, Australia has not yet developed a large-scale evidence base at either state or national levels to demonstrate how the requirements of published professional standards are satisfied (Wyatt-Smith et al., 2017). Further, the disconnect between the academic program and the practical school-based program is widely reported (Daza et al., 2021) and national and international benchmarking remains in its infancy. Arguably, the recommendation with the greatest potential for addressing effectiveness in ITE relates to the introduction of TPAs assessed against an agreed benchmark on program completion.

A strengthened role for standards, evidence, and benchmarking

The adoption of the TEMAG recommendations by the Australian Government (Australian Government Department of Education and Training, 2015) introduced the requirements for a strengthened role for professional standards and evidence as well as "robust assessment of teacher education students" that would provide "schools and families the confidence that graduates are classroom ready" (p. 8). These requirements represent arguably the most significant of those adopted from the TEMAG review as drivers of change in teacher preparation. This was a task delegated to AITSL.

Interestingly, the term teaching performance assessment was not used in the TEMAG review, though there are references to initiatives in the United States as discussed in Chapter 2 of this book. The TEMAG call for 'rigorous assessment of classroom readiness' was carried forward by AITSL into a broad framework of a TPA, defined below. This was presented as a summative assessment required for graduation and to be completed within all ITE programs:

> A teaching performance assessment (TPA) is a tool used to assess the practical skills and knowledge of pre-service teachers. Pre-service teachers collect evidence of practice to complete a TPA in the final year of their initial teacher education program. It is assessed by ITE providers and is a requirement for graduation.
>
> *(AITSL, 2017, para. 1)*

However, the specifications or requirements for a TPA and the necessary trialing and validation processes, including complex standard setting, remained unarticulated beyond broad guidelines for some time. In 2017, universities were invited to form consortia for the purposes of designing, developing, and validating the first TPAs in Australia, including standard setting, moderation, and the generation of preservice teacher exemplars of the completed assessment showing the application of the standard. The initial expectation of university engagement was that TPAs were to be developed by consortia consisting of at least five universities with a lead institution.

From their inception, the official function of a TPA was twofold; first, to enable preservice teachers to demonstrate professional competence or classroom readiness against the graduate level of the professional standards for teachers on completion of their ITE program; and second, to generate evidence of the quality and impact of ITE programs: "A TPA must contribute to the suite of evidence that an ITE provider will collect to demonstrate the impact of their program, including impact on pre-service teacher learning and impact of pre-service teachers on school student learning (Program Standard 6.2 and 6.3)" (AITSL, n.d., p. 4). In both functions, there is a common focus on evidence of teaching performance, with 'quality' of that performance to be assessed against external and officially accepted reference standards (Australian Professional Standards for Teachers; AITSL, 2011, see Box 4.1; National Program Standards, AITSL, 2015, see Box 4.2). TPAs are understood as complex competence assessments (see Chapter 2). They are to be undertaken in situ (classrooms) and at the point of program completion requiring a sustained period of independent classroom teaching. They are not simulations of practice; they require preservice teachers to undertake planning, teaching, assessing, and reflecting on their practice. The expectation is that the profession will recognize TPAs as providing an authentic representation of classroom teaching.

BOX 4.1 THE AUSTRALIAN PROFESSIONAL STANDARDS FOR TEACHERS (APST; AITSL, 2011)

The Australian Professional Standards for Teachers consist of seven standards and 37 focus areas across four career stages (Graduate, Proficient, Highly Accomplished, and Lead). The standards cover professional knowledge (e.g., of content, their students, planning, strategies), professional practice (e.g., to create and maintain stimulating, inclusive, and safe learning environments), and professional engagement (e.g., to progress their own learning). They represent the domain-specific knowledges and skills (competencies) identified as characteristic of the teaching profession in Australia.

Standard 1: Know students and how they learn
Standard 2: Know the content and how to teach it
Standard 3: Plan for and implement effective teaching and learning
Standard 4: Create and maintain supportive and safe learning environments
Standard 5: Assess, provide feedback and report on student learning
Standard 6: Engage in professional learning
Standard 7: Engage professionally with colleagues, parents/carers and the community

BOX 4.2 NATIONAL PROGRAM STANDARDS 1.1, 1.2, AND 1.3 (AITSL, 2015, P. 10)

1.1 Program design and assessment processes identify where each Graduate Teacher Standard is taught, practised and assessed and require that pre-service teachers have demonstrated successful performance against all of the Graduate Teacher Standards prior to graduation.

1.2 Program design and assessment processes require pre-service teachers to have successfully completed a final-year teaching performance assessment prior to graduation that is shown to:

a) be a reflection of classroom teaching practice including the elements of planning, teaching, assessing and reflecting

b) be a valid assessment that clearly assesses the content of the Graduate Teacher Standards

c) have clear, measurable and justifiable achievement criteria that discriminate between meeting and not meeting the Graduate Teacher Standards

d) be a reliable assessment in which there are appropriate processes in place for ensuring consistent scoring between assessors

e). include moderation processes that support consistent decision-making against the achievement criteria.

1.3 Providers identify how their pre-service teachers demonstrate a positive impact on student learning in relation to the assessment requirements in Program Standards 1.1 and 1.2.

TPAs that measure elements such as learning theory, differentiation strategies, and assessment practices in action have been recognized as authentic assessments of teaching (Louden, 2015). These assessments should require preservice teachers to demonstrate the "complexity and multi-dimensionality of teaching, reflect on their teaching and provide detailed explanations and rationales for their plans and decisions" (Renshaw, 2012, p. 14). Louden (2015) identified the Deakin Authentic Teacher Assessment and the Melbourne Clinical Praxis Examination as "starting points for further development" (p. 34) of Australian TPAs. Both were designed to assess profession readiness for teaching, and both were to be undertaken as part of the school-based professional experience. While these acted as starting points for addressing issues of evidence, in 2016 there were no extant TPAs involving multiple universities in Australia.

The introduction of TPAs was a catalyst for attempting to shift the focus from *standards as inputs* used to inform *program development*, to evidence of *assessed standards as outputs* of *program implementation*. The requirement for TPAs was to produce evidence of professional competence with "clear, measurable and justifiable achievement criteria that discriminate between meeting and not meeting the Graduate Teacher Standards" and "moderation processes that support consistent decision-making against the achievement criteria" (see Box 4.2; AITSL, 2015, p. 10). The new direction for developing, assessing, and moderating TPAs in Australia opened the opportunity for cross-university collaboration in assessment design and implementation (see Chapters 6 and 9) and moderation for benchmarking across providers (see Chapter 7).

The requirement for benchmarking performance using a TPA was a significant new juncture in teacher education and assessment in Australia. It led to teacher educators working collaboratively with researchers, policy personnel, and other education experts to validate the instrument and set the pass/fail boundary (meeting and not meeting the standard at the minimum; see Chapter 6). The new requirement was for teacher educators to establish a standard representing a Pass and apply this with demonstrated reliability. To date, several TPAs have been endorsed though this has occurred outside of "a nationally agreed benchmarking framework to confirm the passing standard between different teaching performance assessments" (AITSL, 2015, p. 31).

The absence of an agreed benchmarking framework represents a significant gap in policy and practice in the current Australian ITE landscape. This is ironic noting that AITSL has characterized the framework as giving "confidence these assessments are assessing pre-service teachers' competence against the Graduate Teacher Standards consistently" (AITSL, 2015, p. 31). The lesson that can be learned from the Australian experience is that the move to introduce TPAs as a requirement for establishing teaching competence does not, in and of itself, set a common standard across participating providers. The deeper layers of implementing a TPA concern how teacher education providers are supported to develop knowledge, skills, and expertise – know-how – in how to use the new data and reports of analyses that can be generated from the TPA. Many of the providers in Australia are yet to embrace the understanding of teacher education as evidence-informed practice.

The next section presents a discussion of the underpinning constructs that informed the development of the GTPA®[1]. It characterizes the development process as a shuttling, to and fro, across research, policy, and practice, sharing professional expertise through large-scale collaboration.

Designing the Graduate Teacher Performance Assessment (GTPA)

The policy decision that a TPA was required for graduation from Australian teacher education programs brought assessment of professional competence into the spotlight as never before. There are three features that distinguish Australia's approach to TPA development and implementation in the broader context of ITE policy-driven reform. These are (1) the opportunity in 2016 for universities across the country to work in consortia to conceptualize, design, and trial a new TPA; (2) the recognition of teacher educators as the primary agents of change, including through their role in scoring the assessment; and (3) placing cross-institutional moderation (CIM) as a central feature in quality assurance systems and processes. From the beginning, there was the opportunity for the profession to drive accountability in ways unprecedented in teacher education in this country.

Researchers in the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University, began work in 2016 on the development of the GTPA (https://www.graduatetpa.com). The GTPA is best described as an authentic, complex performance assessment of teaching used to make decisions about licensure for all teacher education graduates. It was designed to generate evidence of the full teaching and assessment cycle and is undertaken in a classroom over a sustained period, as advocated by Mayer (2014). In the next sections, we describe the three lines of investigation that we took to establish the content and construct validity of the GTPA. These involved (1) a first literature review to inform the conceptualization of the instrument and in particular, the content validity of aspects that were core to the demonstration of competence in classroom practice; and (2) a second literature review to focus on the constructs of teaching as a complex practice – the review extended beyond the field of teaching. Drawing on the outcomes of (1) and (2) and the researchers' expertise in teacher education, student learning, and assessment, the instrument and related scoring rubric were designed. The expectation was that the assessment would align closely with the APST, but the approach we took was not to treat the latter as a checklist. Rather, after the assessment was conceptualized and designed, the APST as a set was used as a reference point to check the scope and coverage of expected teaching, learning, and assessment practices. In the third line of inquiry, the assessment was examined by recognized experts in the fields of teaching, education policy, curriculum design, learning, and assessment to establish the extent to which it had credibility with teachers in the field.

The first cycle of the literature review: Conceptualization and content validity

In this book, validity is understood as a unitary concept and is taken to mean: "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" (American Educational Research Association/American Psychological Association/National Council on Measurement in Education [AERA/APA/NCME], 2014, p. 14). This interpretation should be informed by "an analysis of the relationship between the content of a test and the construct it is intended to measure" (AERA/APA/NCME, 2014, p. 14).

In the AITSL-funded 2017 trial of two consortia TPAs in Australia, four practices – planning, teaching, assessing, and reflecting – constituted the framework for validity evidence (AITSL, 2015). These were not elaborated by AITSL. They were however taken as core to the construct of teaching to be assessed and as such carried significant implications for what counted as valued validity evidence of teaching competence. In the following discussion, we address the content and construct representation of the four practices as structural features of the GTPA. In approaching the GTPA as a licensure test, the research team worked from the position that "the major facets that are relevant to the purpose for which the occupation is regulated can be specified, and experts in that occupation can be asked to assign test items to the categories defined by those facets. These or other experts can then judge the representativeness of the chosen items" (AERA/APA/NCME, 2014, p. 14).

A review of the theoretical and research literature was undertaken to investigate aspects of teaching that were core to the demonstration of competence in classroom practice (see Box 4.3 for an abridged summary of the literature review to inform the initial design of the instrument). We initially focused on planning, teaching, assessing, and reflecting, considering that these are represented in TPAs in the United States as relevant precedents (PACT and edTPA, see Chapter 2). However, Accreditation Standard 1.3 (see Box 4.2) called for evidence of how preservice teachers "demonstrate a positive impact on student learning". This informed the research team's decision to include *appraising* as a fifth GTPA practice (Box 4.3). We understood appraising for impact on teaching to involve two interrelated dimensions:

1. An evaluation of the effectiveness of the teaching sequence as implemented for student learning with reference to student work and learning behaviors.
2. An examination of the effectiveness of teaching decisions, made throughout a teaching sequence, in efforts to progress student learning.

Appraising goes beyond reflecting on the effectiveness of one's teaching (Program Standard 1.2). It involves a deliberate shift from what is being taught to what and how students are learning in the classroom. This thinking is evident for a preservice teacher when they demonstrate an openness to 'seeing' student learning, at the whole class and individual levels, and linking this to the part played by their teaching. Through adopting a critical inquiry mindset, preservice teachers continually interrogate issues of evidence: Fitness-for-purpose and what evidence is useful to discern growth in individual learners. Interrogating teaching through a laser-like focus on student learning was understood as the practice of 'appraising' and 'demonstrating impact' of teaching.

Several authors have written about appraisal in terms of student actions. For example, Black and Wiliam (1998) identified appraisal as the actions of classroom students to recognize gaps and then plan and carry out necessary actions to address the gaps. Sadler (2010), in discussing students' use of feedback in higher education, used judgment and appraisal as synonyms, thus students made judgments about (appraised) the quality of their work to attend to those sections which required further attention. Understanding appraisal as part of the continuous learning for a teacher has been shown to be an important part of professional practice (Organization for Economic Cooperation and Development [OECD], 2013). Grimaldi (2012) discussed appraisal as an aspect of strategic action "where actors actualize some of the possibilities opened by the interweaving of discourse and social structures through their selectivities" (p. 455). Grimaldi is referring to the privileging of some ways of doing and being (e.g., by choosing a pedagogic strategy, or adopting a particular professional identity) in response to context and the discourses of that specific context. Conducting a critical appraisal of a range of data and evidence, selecting that evidence pertinent to a situation, and justifying decisions and actions based on the combination of evidence has been described in the literature as evidence-informed practice (e.g., Mandinach, 2012; Van der Kleij et al., 2015).

While these studies varied in their purposes and methodologies, common to them all is an interest in appraisal as a complex activity where plans are not necessarily enacted as intended. Thus, appraisal is understood as critical inquiry (Serafini, 2000; Wyatt-Smith & Bridges, 2006; Wyatt-Smith & Gunn, 2010) grounded in context. Teachers' use of assessment as critical inquiry is related to "the interactivity of curriculum, pedagogy and assessment as foundational elements for quality learning" (Wyatt-Smith & Bridges, 2006, p. 13). This involves a professional mindset that views assessment as an ongoing aspect of professional work, as teachers continuously appraise their practice and adjust their teaching and ways of interacting in response to their observations and other collected data. Serafini (2000) describes assessment as an inquiry paradigm, involving teachers in using "various qualitative and quantitative assessment techniques to inquire about particular learners and their learning processes. It is a process of inquiry, and a process of interpretation, used to promote reflection concerning students' understandings, attitudes, and… abilities" (p. 387). She goes on to say that "within this paradigm, the purpose of assessments is a deeper understanding of individual learners in their specific learning contexts" (p. 387).

BOX 4.3 ABRIDGED SUMMARY OF THE LITERATURE REVIEW (2014–2017) AS INFORMING THE CONCEPTUALIZATION AND DESIGN OF THE INSTRUMENT

The literature review examined each of the practices in turn, guided by a research question.
Practice 1: Planning – What evidence would indicate that preservice teachers have mastered the core skills of planning? Included in the review was literature that examined:

1. The planning process (e.g., Norman, 2011); planning as front-ending assessment (Wyatt-Smith & Bridges, 2006);
2. The alignment of curriculum, assessment, and pedagogy (e.g., Klenowski & Wyatt-Smith, 2013);
3. The use of a range of data sources to inform planning as an ongoing pedagogic practice (e.g., Bailey & Drummond, 2006; Mandinach, 2012);
4. Planning for a range of core competencies such as literacy, numeracy, and technology (e.g., Bitter & Legacy, 2008; Czislowski-McKenna et al., 2006; Geiger et al., 2015a; Geiger et al., 2015b; Koh et al., 2017);
5. Selecting teaching strategies that best promote inclusion as well as each child's thinking, creativity, and problem-solving skills (e.g., Dweck, 2000; Epstein, 2006; Kalantzis & Cope, 2012; Lucas, 2016; Marzano & Heflebower, 2012; Puntambekar & Du Boulay, 1997; Robinson & Aronica, 2015; Tomlinson et al., 2003).

Practice 2: Teaching – What are the different conceptions of teaching? Included in the review was literature that examined:

1. Conceptions of teacher knowledge and learning (e.g., Connell, 2009; Hollins, 2011; Shulman, 1986);
2. Effective teaching strategies (e.g., Schoenfeld, 2014; Shulman, 1987);
3. Teaching a range of core competencies such as literacy, numeracy, and technology (e.g., Bitter & Legacy, 2008; Czislowski-McKenna et al., 2006; Geiger et al., 2015a; Geiger et al., 2015b; Gunn & Wyatt-Smith, 2011; Schmidt et al., 2009; Tatto et al., 2012; Zhang et al., 2015);
4. Differentiated practice and culturally relevant pedagogy (e.g., Banks et al., 2005; Schmeichel, 2012; Tomlinson et al., 2003).

Practice 3: Assessing – What are effective assessment practices? Included in the review was literature that examined:

1. Assessment processes (e.g. Stiggins et al., 2004; Stobart, 2008);

2. The relationship between assessment and learning (e.g., Earl, 2003; Gardner, 2012; Reeves, 2007);
3. Formative assessment including feedback (e.g., Black & Wiliam, 1998; Hattie & Timperley, 2007);
4. Developing students' evaluative expertise, goal-setting, and tracking learning (e.g., Andrade & Brown, 2016; Broadfoot, 2007; Costa & Kallick, 2004; Sadler, 1989; Stobart, 2014);
5. Summative assessment (e.g. Harlen, 2005; Moss, 2013; Pellegrino et al., 2001), including assessing a range of core competencies such as literacy, numeracy, and technology (Brookhart, 2010; Kimber & Wyatt-Smith, 2014; Wyatt-Smith & Cumming, 2003);
6. Differentiated assessment (e.g., Moon, 2016; Tomlinson & Moon, 2013);
7. Judgment-making (e.g., Brookhart, 2013; Klenowski & Wyatt-Smith, 2013);
8. Moderation (e.g., Adie & Klenowski, 2016; Wyatt-Smith & Colbert, 2014).

Practice 4: Reflecting – How has reflecting been conceptualized? Included in the review was literature that examined:

1. Reflection-in-action (e.g., Schön, 1987);
2. Reflection as a cycle (e.g., Korthagen & Vasalos, 2005);
3. Reflection and professional identity (e.g., Beijaard et al., 2004);
4. Reflection in teacher education (e.g., Adie & Tangen, 2015; Gelfuso, 2017; Jay & Johnson, 2002; Jones & Charteris, 2017).

Practice 5: Appraising – How can the quality and impact of teaching be evidenced? Included in the review was literature that examined:

1. Developing preservice teachers' skills in justification and evidence-based explanation (e.g., Black & Wiliam, 1998; Grimaldi, 2012; Sadler, 2010);
2. Assessment as critical inquiry (e.g., Serafini, 2000; Wyatt-Smith & Bridges, 2006; Wyatt-Smith & Gunn, 2010);
3. The use of assessment to inform future teaching and learning (e.g., Adie et al., 2013; Black et al., 2011);
4. Teaching as evidence-informed practice (e.g., Mandinach, 2012; Van der Kleij et al., 2015).

The second cycle of the literature review: Constructs of teaching

As previously mentioned, a key focus in designing the GTPA was identifying the characteristics of a final-year summative assessment that would provide evidence of teaching competence. A related focus was how this type of assessment was different from other assessments required throughout teacher education, including the academic component and the school-based practical component. This was of particular significance since the assessment served a summative purpose and was therefore high-stakes: It was to act as a gatekeeper for graduation and licensure for entering the teaching profession. The challenge was to design an assessment that captures performance-in-context that can be measured reliably and is applicable across contexts. This required demonstrating reasoning and problem-solving in ways that bring together the theoretical and practical contextual knowledges to show the process of decision-making in teaching.

In addressing these aspects of practice, we investigated Miller's (1990) framework for clinical assessment in the field of medicine. This framework, presented as a pyramid, differentiates four areas for clinical assessment: *knows* (knowledge), *knows how* (competence), *shows how* (performance), and *does* (action). Miller (1990) asserted that "Tests of knowledge are surely important, but they are also incomplete tools … if we really believe there is more to the practice of medicine than knowing" (p. S63).

The GTPA development team was of the view that there was more to the practice of teaching than knowing. Also consistent with Miller's characterization of performance in medicine, is that in teacher preparation, preservice teachers "must develop… the skill of acquiring information from a range of human and laboratory [school] sources, to analyze and interpret these data, and finally to translate such findings into a rational diagnostic or management plan. It is this quality of being functionally adequate, or of having sufficient knowledge, judgment, skill, or strength for a particular duty that Webster defines as competence" (p. S63).

Referring to medicine, Miller (1990) identified "a growing body of evidence suggesting that… judgments are generally based on limited observation and equally limited sampling of clinical problems (which means an inadequate data base)" (p. S63). He went on to assert that such judgments "seem more often related to the product of student interaction with patients, that is, to the accuracy of diagnosis and the nature of management than through the process through which these conclusions were reached" (p. S63). These observations came to be influential in our thinking about what counted as demonstrations of competence for teaching.

The thinking processes about student learning, responsiveness to teaching, and what was required in next-step teaching were explored in developing the GTPA. Informed by Miller (1990), it was evident that insights into teaching performance could not be obtained through demonstrations of actions alone. Our interest was in how competence assessment in teaching could make decision processes visible, including allowing demonstrations of teacher uncertainty and teacher learning during practice. A few years earlier, Shulman (1986) investigated the knowledge base of teachers and the relationship between their content knowledge and pedagogic decisions. He described pedagogical content knowledge (PCK) as including teacher knowledge of "the most useful forms of representation…, the most powerful analogies, illustrations, examples, explanations, and demonstrations—in a word, the ways of representing and formulating the subject that make it comprehensible to others" (p. 9). Effective teachers choose, from a range of strategies, those that fit the context and specific needs of a student to progress their learning of the subject matter. Drawing on Shulman's (1986) work, a competence assessment of teaching would take account of the content being taught and student level of understanding, in justifying chosen strategies.

The next point of focus was to investigate how we would include the moment-by-moment decision-making (Schoenfeld, 2014) in the GTPA in the course of real-time teaching and observing learning and learners. We took this decision-making to involve intentional practice through reflexivity. Applied to teaching, this involves reflection plus action. It includes appraising evidence of learning, identifying possibilities for improving practice, discerning next steps for teaching and learning, and making decisions to enact new practice with the intention of improving student learning.

Schoenfeld (2014) posited the complexity of teaching as a craft that stems in part from situational variables, including the interactions that occur between teachers and students, and students and students. These are characteristically dynamic and remain unpredictable. Added to this are the opportunities for intertwining research and practice to inform "classroom decision making on a moment-by-moment basis" (p. 405). In the GTPA, assessment invites preservice teachers to make explicit their reasoning and decision-making in five practices informed by relevant theory and research (see Box 4.4).

BOX 4.4 THE FIVE CORE PRACTICES THAT CONSTITUTE THE GTPA

Practice 1. *Planning* using data: Collecting and interpreting a range of data to establish students' learning needs and current levels of performance, and to inform planning and teaching; and the alignment of curriculum, assessment, and pedagogy with a focus on learning.

Practice 2. *Teaching* and learning: Employing a range of suitably challenging and engaging teaching and learning strategies to meet the diverse needs of students.

Practice 3. *Assessing*, feedback, and professional judgment: Selecting and using a variety of assessment tools and practices to provide feedback, make judgments, moderate grades, and inform next-step teaching.

Practice 4. *Reflecting* on teaching: Analyzing the scope and sufficiency of initial and ongoing data choices, differences between intended and enacted practice, and decisions for future teaching supported through relevant research and theory.

Practice 5. *Appraising* the impact of teaching: Evaluating the effectiveness of teaching and demonstrating its impact on student learning in relation to the chosen actions to progress student learning.

Following the outcomes of various feedback loops and inquiry into the GTPA, the next step was to develop a detailed description of the five practices that constituted the conceptual framework for the assessment. This required identifying the knowledge, skills, abilities, competencies, processes, and characteristics to be assessed. Consistent with the standards for educational and psychological testing (AERA/APA/NCME, 2014), "the framework indicates how the construct as represented is to be distinguished from other constructs and how it should relate to other variables" (p. 11). The constructs in the GTPA are understood as distinguishable but are not taken up as a linear sequence of activities.

In completing the assessment, preservice teachers are required to identify and collect initial data to inform their planning for practice and to establish their students' entry-level knowledge, skills, and capabilities prior to beginning teaching. They are also required to show how a range of data is collected continuously through their teaching and used to adjust their planning in response to student learning needs. Formative assessments of student work, including feedback, contribute to this evidence of learning and further inform adjustments to planning and teaching. Assessment activities, including the review of student work samples and related judgments, contribute to preservice teachers' reflections on teaching practice as they check their pedagogic decisions and review prior assumptions about learners and learners' readiness to proceed. This informs how they appraise their teaching and discern its effectiveness to support student learning.

Further, in demonstrating their teaching, preservice teachers draw together research literature and theory, and various policy documents and materials including school policy, classroom context, and available student record data. Using these data, preservice teachers give an account of their work and justify their decisions and actions drawing on cases where learning improved and where no improvement was evident. A critical design feature of the GTPA was to capture thinking about practice as an explanation of professional judgment and decision-making in teaching. This also included thinking about future actions – next-step teaching – to progress student learning. In these ways, the design of the GTPA aimed to enable preservice teachers to undertake the principled collection of authentic and valid evidence of their teaching performance.

Standards and the design of the GTPA scoring rubric

The GTPA scoring rubric was designed to align with the core practices of *Planning*, *Teaching*, *Assessing*, *Reflecting*, and *Appraising* in response to the Graduate Teacher Standards (AITSL, 2011) and the Program Standards (AITSL, 2015) as previously described. While we recognized that a single piece of evidence, such as the GTPA, can address multiple descriptors across the seven Graduate Standards, we also recognized that context is likely to impact the opportunity to demonstrate competence in some of the standards. This meant that what could be authentically assessed in one context in which the GTPA is completed may not be appropriate in another context. The design of the rubric needed to be relevant to all contexts. This led to the decision that standard descriptors that were context-dependent or broader than the focus on student learning were not mandatory inclusions in the performance assessment. For example, Graduate Standard 7.3 – Engage with the parents/carers – required preservice teachers to "Understand strategies for working effectively, sensitively, and confidentially with parents/carers" (AITSL, 2011, p. 19). While preservice teachers need to learn how to work with the parents of their students, it cannot be guaranteed in all professional practice contexts that this opportunity would be provided. A decision was made that the GTPA did not require an explicit response to this standard, and that it would not be part of the criteria descriptors in the rubric. This standard would be covered in other aspects of the ITE program.

To determine what could be authentically included in the rubric, several characteristics needed to be considered. These included: Phase of schooling (primary, secondary, early years), focus class or student group, the selection of three 'focus students' chosen as representative of a range of achievement levels in the class and selected for focused analysis, the role and expectations of the supervising teacher, and any specific requirements added by an ITE provider as a focus for a particular cohort of preservice teachers.

The rubric was designed so that preservice teachers had multiple opportunities to provide evidence of competence in the selected Graduate Standards (see Table 4.2). For example, evidence of Graduate Standard 5.4 – *Interpret student data: Demonstrate the capacity to interpret student assessment data to evaluate student learning and modify teaching practice* – could be found in the criteria:

- *Planning* as preservice teachers collected and used data from a range of sources to inform planning.
- *Teaching* as data are collected and used to modify plans and teaching strategies, including in-the-moment teaching decisions.
- *Assessing* as data are analyzed to inform judgment decisions.
- *Reflecting* as preservice teachers consider their data-informed pedagogic decisions, alternate actions, and possible future directions for teaching.
- *Appraising* as they consider their collected data to show the impact of their teaching on student learning.

Expert review

As mentioned earlier, the GTPA was subject to systematic review during its development by teaching experts and researchers in the field of learning and assessment who made "expert judgments of the relationship between parts of the test and the construct" (AERA/APA/NCME, 2014, p. 14). Experts included members of AITSL; teacher regulatory authorities, in particular, Queensland College ofTeachers (QCT); the GTPA Steering Group with representatives from 11 peak national bodies; the QCT Principals' Engagement Reference Group (PERG) comprised of principals from different phases of schooling and sectors; and senior officers of the Queensland Teachers Union and the Independent Education Union – Queensland

TABLE 4.2 Illustrative mapping of selected professional standards (APST; AITSL, 2011) and numbered descriptors in the GTPA rubric

APST (graduate) descriptors	*GTPA numbered descriptors in stated criteria for the required practices (for evidence of 'assessed')*
1.3 Demonstrate knowledge of teaching strategies that are responsive to the learning strengths and needs of students from diverse linguistic, cultural, religious and socioeconomic backgrounds	2. Establish students' current level of performance, desired level of performance, and readiness for learning 5. Employ a range of suitably challenging and engaging teaching and learning strategies that connect to and build on students' prior learning **6. Provide differentiated teaching and learning opportunities** 8. Make suitable adjustments to teaching based on ongoing student data gathering and analysis 9. Select and use a variety of assessment tools and practices, addressing fitness for purpose and principles of inclusion 10. Provide feedback to learners to inform student self-assessment, goal setting, and to progress learning
2.3 Use curriculum, assessment and reporting knowledge to design learning sequences and lesson plans	1. Collect, interpret and use a variety of student data and evidence for diagnostic, formative and summative purposes 3. Use the official curriculum and other relevant materials to plan connected teaching and learning sequences **6. Provide differentiated teaching and learning opportunities** 7. Teach the general capabilities, including literacy and numeracy, required for student success in learning 8. Make suitable adjustments to teaching based on ongoing student data gathering and analysis 9. Select and use a variety of assessment tools and practices, addressing fitness for purpose and principles of inclusion 11. Make judgements of the quality of student work with reference to curriculum and achievement standards 12. Engage in moderation of student work 14. Identify and describe differences between planned and enacted teaching, and related pedagogical reasoning 15. Discuss how evidence of learning was used to monitor student progress and to modify teaching and assessment strategies 16. Identify and justify future teaching and assessment practices in relation to relevant theory

(*Continued*)

TABLE 4.2 (Continued)

APST (graduate) descriptors	*GTPA numbered descriptors in stated criteria for the required practices (for evidence of 'assessed')*
5.4 Demonstrate the capacity to interpret student assessment data to evaluate student learning and modify teaching practice	1. Collect, interpret and use a variety of student data and evidence for diagnostic, formative and summative purposes 2. Establish students' current level of performance, desired level of performance, and readiness for learning 5. Employ a range of suitably challenging and engaging teaching and learning strategies that connect to and build on students' prior learning **6. Provide differentiated teaching and learning opportunities** 8. Make suitable adjustments to teaching based on ongoing student data gathering and analysis 9. Select and use a variety of assessment tools and practices, addressing fitness for purpose and principles of inclusion 11. Make judgements of the quality of student work with reference to curriculum and achievement standards 12. Engage in moderation of student work 13. Describe and analyse the scope and sufficiency of initial and ongoing data choices for identifying students' learning needs and informing next-step teaching 15. Discuss how evidence of learning was used to monitor student progress and to modify teaching and assessment strategies 16. Identify and justify future teaching and assessment practices in relation to relevant theory 17. Connect theory, enacted practice and the curated body of evidence to evaluate the effectiveness of teaching, and demonstrate its impact on student learning 18. Examine and discuss how teaching decisions were effective or not effective in progressing student learning and why

and Northern Territory. Feedback from the reviews informed the collection and use of validity evidence for the design and redesign of the instrument. Experts were asked: Will the assessment produce evidence of what you would expect a beginning teacher to know and be able to do?

This large-scale collaboration also involved two rounds of audit of the achievement criteria against the APST Graduate level and the requirements of Program Standard 1.2 undertaken by the QCT and school-based experts. The initial decision was made for researchers and policy personnel to work independently of each other in the early design work. This decision was informed by the insight that different stakeholders bring expertise from their domains of practice that would provide a range of perspectives for seeing and reseeing the instrument. It also considered the importance of how criteria and standards are conceptualized and understood, noting the concern raised by Furlong (2015), as mentioned in Chapter 1, that standards can be used as a checklist of unrelated elements having a narrowing influence on learning and teaching. The teams of experts were asked to analyze draft versions of the GTPA to discern:

- Those stated criteria requiring demonstration of knowledge and understanding and those requiring demonstration through action.
- The relationship of the criteria to the five practices to be assessed.
- Fitness of the criteria for the purpose of assessing teaching in teacher preparation programs for the phases of schooling.
- Equity of opportunity for preservice teachers in their school-based practical program to demonstrate the aspects of practice to be assessed using the criteria.

These processes were undertaken over a two-year review period (2016–2017), following which stakeholders recognized the GTPA as an authentic assessment that reflected the complex and interrelated facets of professional practice in teaching. Table 4.2 provides an example of three selected descriptors of professional standards mapped against the descriptors of expected performance in the GTPA stated criteria. The mapping was undertaken as an audit to trace opportunities for providing evidence of competence assessed against the professional standards. Table 4.2 presents the GTPA descriptors as numbered to show repeat opportunities to demonstrate the selected standard. For example, in the table, the GTPA descriptor 'Provide differentiated teaching and learning opportunities' is bolded to illustrate this point.

Conclusion

This chapter gives readers an insight into an Australian case of designing a TPA, with subsequent chapters exploring its development. The discussion focused on the design of the GTPA as a single, culminating, or summative assessment undertaken during the final year of a teacher education program. However, the design of a TPA is the tip of the iceberg. The value of a TPA lies in its transformative potential for realizing change in teacher education. The main means to achieve this is through collaborative inquiry into practice. Here, we offer readers nine questions relevant to a decision to introduce and use a TPA intended to assess preparedness to teach. Together they shine a light on what a TPA can contribute as a summative assessment and beyond this to realizing improvement in ITE.

1. What is the conceptualization and design of the instrument? What are its underpinning constructs? What evidence is available to show it is fit-for-purpose?
2. What is known about how the assessment was validated? What was the approach to standard setting?
3. What is the expected approach to scoring and determining rater and inter-rater reliability?
4. What are the conditions under which the assessment can be implemented to ensure assessment integrity and fidelity of implementation?
5. What evidence is available regarding the application of a common standard across sites where the assessment is used? What benchmarking, if any, is undertaken?
6. How is evidence produced by the assessment analyzed?
7. How are reports of performance quality and scoring outcomes and trends over time used to inform curriculum review and program renewal?
8. Are there any barriers in the assessment design or implementation practices that could limit opportunities for success for students from diverse cultural, linguistic, and socio-economic backgrounds?
9. What longitudinal studies could be undertaken to investigate the impact of the assessment on graduate quality and the effectiveness of classroom practice?

The following chapters explore these questions through our lens of the GTPA and the openings for innovation and collaboration that it called forth.

Note

[1] Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

References

Adie, L., & Klenowski, V. (2016). Moderation and assessment. In *Encyclopedia of Educational Philosophy and Theory*. Springer. http://link.springer.com/10.1007/978-981-287-532-7_393-1

Adie, L., Lloyd, M., & Beutel, D. (2013). Identifying discourses of moderation in higher education. *Assessment and Evaluation in Higher Education*, *38*(8), 968–977. https://doi.org/10.1080/02602938.2013.769200

Adie, L., & Tangen, D. (2015). The use of multimodal technologies to enhance reflective writing in teacher education. In M. E. Ryan (Ed.), *Teaching reflective learning in higher education: A systematic approach using pedagogic patterns* (pp. 127–138). Springer. https://doi.org/10.1007/978-3-319-09271-3_9

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andrade, H., & Brown, G. (2016). Student self-assessment in the classroom. In G. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 319–334). Routledge.

Australian Government. (2015). *Teacher Education Ministerial Advisory Group. Action now: Classroom ready teachers – Australian government response*. https://docs.education.gov.au/documents/australian-government-response-action-now-classroom-ready-teachers-report

Australian Institute for Teaching and School Leadership (AITSL). (2011). *Australian professional standards for teachers*. https://www.aitsl.edu.au/teach/standards

Australian Institute for Teaching and School Leadership (AITSL). (2015). *Accreditation of initial teacher education programs in Australia: Standards and procedures*. https://www.aitsl.edu.au/docs/default-source/national-policy-framework/accreditation-of-initial-teacher-education-programs-in-australia.pdf?sfvrsn=e87cff3c_28

Australian Institute for Teaching and School Leadership (AITSL). (2017). *Teaching performance assessment*. https://www.aitsl.edu.au/deliver-ite-programs/teaching-performance-assessment

Australian Institute for Teaching and School Leadership (AITSL). (n.d.). *Teaching performance assessment: Program Standard 1.2*. https://www.aitsl.edu.au/docs/default-source/initial-teacher-education-resources/tpa/tpa-fact-sheet.pdf?sfvrsn=1410cb3c_6

Bailey, A. L., & Drummond, K. V. (2006). Who is at risk and why? Teachers' reasons for concern and their understanding and assessment of early literacy. *Educational Assessment*, *11*(3–4), 149–178.

Banks, J., Cochran-Smith, M., Moll, L., Richert, A., Zeichner, K., LePage, P., Darling-Hammond, L., Duffy, H., & McDonald, M. (2005). Teaching diverse learners. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 232–274). Jossey-Bass.

Beijaard, D., Meijer, P. C., & Verloop, N. (2004). Reconsidering research on teachers' professional identity. *Teaching and Teacher Education*, *20*(2), 107–128. https://doi.org/10.1016/j.tate.2003.07.001

Bitter, G. G., & Legacy, J. M. (2008). *Using technology in the classroom* (7th Ed.). Pearson.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice*, *18*(4), 451–469. https://doi.org/10.1080/0969594X.2011.557020

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Broadfoot, P. (2007). *An introduction to assessment*. Continuum.

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Brookhart, S. M. (2013). Grading. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 257–271). SAGE.

Cochran-Smith, M., Piazza, P., & Power, C. (2013). The politics of accountability: Assessing teacher education in the United States. *Educational Forum*, 77(1), 6–27. https://doi.org/10.1080/00131725.2013.739015

Connell, R. (2009). Good teachers on dangerous ground: Towards a new view of teacher quality and professionalism. *Critical Studies in Education 50*(3), 213–229. https://doi.org/10.1080/17508480902998421

Costa, A. L., & Kallick, B. (2004). *Assessment strategies for self-directed learning*. Corwin Press.

Craven, G., Beswick, K., Fleming, J., Fletcher, T., Green, M., Jensen, B., Leinonen, E., & Rickards, F. (2014). *Action now: Classroom ready teachers*. Teacher Education Ministerial Advisory Group, TEMAG. Department of Education. Australia. Retrieved on 9 April 2019 from: https://docs.education.gov.au/documents/action-now-classroom-ready-teachers-report

Czislowski-McKenna, A. T., Cumming, J. J., Wyatt-Smith, C. M., & Elkin, J. (2006). *Literacy teaching and learning in Victorian schools. Paper no. 9*. Department of Education and Training.

Daza, V., Gudmundsdottir, G. B., & Lund, A. (2021). Partnerships as third spaces for professional practice in initial teacher education: A scoping review. *Teaching and Teacher Education*, *102*, 1–14. https://doi.org/10.1016/j.tate.2021.103338

Duckor, B., Castellano, K. E., Téllez, K., Wihardini, D., & Wilson, M. (2014). Examining the internal structure evidence for the performance assessment for California teachers: A validation study of the elementary literacy teaching event for Tier I teacher licensure. *Journal of Teacher Education*, *65*(5), 402–420. https://doi.org/10.1177/0022487114542517

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Psychology Press.

Earl, L. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Corwin Press.

Epstein, A. S. (2006). *The intentional teacher: Choosing the best strategies for young children's learning*. National Association for the Education of Young Children.

Furlong, J. (2015). *Teaching tomorrow's teachers: Options for the future of initial teacher education in Wales*. https://gov.wales/sites/default/files/publications/2018-03/teaching-tomorrow%E2%80%99s-teachers.pdf

Gardner, J. (Ed.). (2012). *Assessment and learning* (2nd Ed.). SAGE.

Geiger, V., Forgasz, H., & Goos, M. (2015a). A critical orientation to numeracy across the curriculum. *ZDM*, *47*(4), 611–624. https://doi.org/10.1007/s11858-014-0648-1

Geiger, V., Goos, M., & Dole, S. (2015b). The role of digital technologies in numeracy teaching and learning. *International Journal of Science and Mathematics Education*, *13*(5), 1115–1137. https://doi.org/10.1007/s10763-014-9530-4

Gelfuso, A. (2017). Facilitating the development of preservice teachers' pedagogical content knowledge of literacy and agentic identities: Examining a teacher educator's intentional language choices during video-mediated reflection. *Teaching and Teacher Education*, *66*, 33–46. https://doi.org/10.1016/j.tate.2017.03.012

Grimaldi, E. (2012). Analysing policy in the context(s) of practice: A theoretical puzzle. *Journal of Education Policy*, *27*(4), 445–465. https://doi.org/10.1080/02680939.2011.647926

Gunn, S., & Wyatt-Smith, C. (2011). Learning difficulties, literacy and numeracy: Conversations across the fields. In C. Wyatt-Smith, J. Elkins, & S. Gunn (Eds.), *Multiple perspectives on difficulties in literacy and numeracy learning* (pp. 17–48). Springer.

Harlen, W. (2005). Teachers' summative practices and assessment for learning: Tensions and synergies. *The Curriculum Journal*, *16*(2), 207–223. https://doi.org/10.1080/09585170500136093

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. https://doi.org/10.3102/003465430298487

Hollins, E. (2011). Teacher preparation for quality teaching. *Journal of Teacher Education*, *62*(4), 395–407. https://doi.org/10.1177/0022487111409415

Jay, J., & Johnson, K. (2002). Capturing complexity: A typology of reflective practice for teacher education. *Teaching and Teacher Education*, *18*, 73–85. https://doi.org/10.1016/S0742-051X(01)00051-8

Jones, M., & Charteris, J. (2017). Transformative professional learning: An ecological approach to agency through critical reflection. *Reflective Practice*, *18*(4), 496–513. https://doi.org/10.1080/14623943.2017.1307729

Kalantzis, M., & Cope, B. (2012). *New learning: Elements of a science of education* (2nd Ed.). Cambridge University Press.

Kimber, K., & Wyatt-Smith, C. (2014). Designing next generation assessment. In C. Wyatt-Smith, V. Klenowski, & P. Colbert. (Eds.), *Designing assessment for quality learning* (pp. 357–371). Springer.

Klenowski, V., & Wyatt-Smith, C. (2013). *Assessment for education: Standards, judgement and moderation*. SAGE.

Koh, J. H. L., Chai, C. S., & Lim, W. Y. (2017). Teacher professional development for TPACK-21CL. *Journal of Educational Computing Research*, *55*(2), 172–196. https://doi.org/10.1177/0735633116656848

Korthagen, F., & Vasalos, A. (2005). Levels in reflection: Core reflection as a means to enhance professional growth. *Teachers and Teaching: Theory and Practice*, *11*(1), 47–71. https://doi.org/10.1080/1354060042000337093

Louden, W. (2015). *Standardised assessment of initial teacher education: Environmental scan and case studies*. Australian Institute for Teaching and School Leadership (AITSL).

Lucas, B. (2016). A five-dimensional model of creativity and its assessment in schools. *Applied Measurement in Education*, *29*(4), 278–290. http://dx.doi.org/10.1080/08957347.2016.1209206

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*(2), 71–85. https://doi.org/10.1080/00461520.2012.667064

Marzano, R. J., & Heflebower, T. (2012). *Teaching and assessing 21st century skills*. Marzano Research Laboratory.

Mayer, D. (2014). Forty years of teacher education in Australia: 1974–2014. *Journal of Education for Teaching*, *40*(5), 461–473. https://doi.org/10.1080/02607476.2014.956536

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine 65*(9), S63–S67. https://doi.org/10.1097/00001888-199009000-00045

Mills, M., & Goos, M. (2017). The place of research in teacher education? An analysis of the Australian Teacher Education Ministerial Advisory Group Report *Action Now: Classroom Ready Teachers*. In M. A. Peters, B. Cowie, & I. Menter (Eds.), *A companion to research in teacher education* (pp. 637–650). Springer.

Moon, T. R. (2016). Differentiated instruction and assessment: An approach to classroom assessment in conditions of student diversity. In G. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 284–301). Routledge.

Moss, C. M. (2013). Research on classroom summative assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 235–250). SAGE.

Norman, P. (2011). Planning for what kind of teaching? Supporting cooperating teachers as teachers of planning. *Teacher Education Quarterly*, *38*(3), 49–68.

Organization for Economic Cooperation and Development (OECD). (2013). *Teachers for the 21st century: Using evaluation to improve teaching*. OECD Publishing.

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education*, *57*(1), 22–36. https://doi.org/10.1177/0022487105284045

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press.

Puntambekar, S., & Du Boulay, B. (1997). Design and development of MIST: A system to help students develop metacognition. *Journal of Educational Computing Research*, *16*(1), 1–35.

Reeves, D. (Ed.). (2007). *Ahead of the curve: The power of assessment to transform teaching and learning*. Solution Tree Press.

Renshaw, P. (2012). *Literature review and environmental scan: Supervising professional experience students*. Australian Institute for Teaching and School Leadership (AITSL). https://apo.org.au/sites/default/files/resource-files/2012/06/apo-nid42118-1181906.pdf

Robinson, K., & Aronica, L. (2015). *Creative schools: Revolutionizing education from the ground up*. Allen Lane.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Sadler, D. R. (2010). Fidelity as a precondition for integrity in grading academic achievement. *Assessment & Evaluation in Higher Education*, *35*(6), 727–743. https://doi.org/10.1080/02602930902977756

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of Teacher Education*, *65*(5), 421–434. https://doi.org/10.1177/0022487114542518

Schmeichel, M. (2012). Good teaching? An examination of culturally relevant pedagogy as an equity practice. *Journal of Curriculum Studies*, *42*(2), 211–231. https://doi.org/10.1080/00220272.2011.591434

Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., & Shin, T. S. (2009). Technological pedagogical content knowledge (TPACK). *Journal of Research on Technology in Education*, *42*(2), 123–149. https://doi.org/10.1080/15391523.2009.10782544

Schoenfeld, A. (2014). What makes for powerful classrooms, and how can we support teachers in creating them? A story of research and practice, productively intertwined. *Educational Researcher*, *43*(8), 404–412. https://doi.org/10.3102/0013189X14554450

Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. Jossey-Bass.

Serafini, F. (2000). Three paradigms of assessment: Measurement, procedure, and inquiry. *Reading Teacher*, *54*(4), 384–393.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–22. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Stiggins, R. J., Arter, J. A., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning: Doing it right – using it well*. Assessment Training Institute.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. Routledge.

Stobart, G. (2014). *The expert learner: Challenging the myth of ability*. Open University Press.

Tatto, M., Schwille, J., Senk, S., Ingvarson, L., Rowley, G., Peck, R., … Reckase, M. (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)*. International Association for the Evaluation of Educational Achievement (IEA) https://files.eric.ed.gov/fulltext/ED542380.pdf

Tomlinson, C., Brighton, C., Hertberg, H., Callahan, C., Moon, T., Brimijoin, K., … Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classroom settings: A review of literature. *Journal for the Education of the Gifted*, *27*(2/3), 119–145. https://doi.org/10.1177/016235320302700203

Tomlinson, C. A., & Moon, T. R. (2013). Differentiation and classroom assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 415–430). SAGE.

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 324–343. https://doi.org/10.1080/0969594X.2014.999024

Wyatt-Smith, C., Alexander, C., Fishburn, D., & McMahon, P. (2017). Standards of practice to standards of evidence: Developing assessment capable teachers. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 250–270. https://doi.org/10.1080/0969594X.2016.1228603

Wyatt-Smith, C., & Bridges, S. (2006). Assessment for learning: An Australian study in middle schooling. In *International association for educational assessment 32nd annual conference*. www.iaea2006.seab.gov.sg/conference/programme.html

Wyatt-Smith, C., & Colbert, P. (2014). An account of the inner workings of standards, judgement and moderation: A previously untold evidence-based narrative. *Informing paper for the review of Queensland Senior Assessment and School Reporting and Tertiary Entrance Processes undertaken by Australian Council for Educational Research (ACER)*. http://www.acer.edu.au/files/Wyatt-SmithColbert_InformingPaper_Final.pdf

Wyatt-Smith, C. M., & Cumming, J. J. (2003). Curriculum literacies: Expanding domains of assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(1), 47–59. https://doi.org/10.1080/09695940301690

Wyatt-Smith, C. M., & Gunn, S. (2010). Towards theorising assessment as critical inquiry. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 83–102). Springer.

Zhang, M., Trussell, R. P., Gallegos, B., & Asam, R. R. (2015). Using math apps for improving student learning: An exploratory study in an inclusive fourth grade classroom. *TechTrends*, *59*(2), 32–39. https://doi.org/10.1007/s11528-015-0837-y

5

TEACHING PERFORMANCE ASSESSMENTS AND CONSIDERATIONS OF FAIRNESS

Joy Cumming and Diana Pullin

Introduction

Australian initiatives to seek 'quality assurance' in universities to improve the education of future teachers (Australian Government Department of Education and Training, 2016) are consistent with directions in many nations (Akiba, 2017; Burns & McIntyre, 2017; Darling-Hammond et al., 2017). As in Australia, many of these initiatives involve setting professional standards for teaching and creating performance assessments to measure teacher competence. These efforts seek a broader reform of the teaching workforce and offer potential opportunities and considerable challenges for the overall improvement of educational outcomes for all students served in an education system.

Australia provides an interesting case study into such initiatives and efforts through the capacity to have national oversight of areas that are simultaneously the responsibility of states and territories. As discussed in Chapter 1, the establishment in 2010 of the Australian Institute for Teaching and School Leadership (AITSL), with the development of the Australian Professional Standards for Teachers (APST; AITSL, 2011; revised 2018) to provide expectations for quality teaching at different professional levels, is one such initiative. The establishment in 2014 of the Teacher Education Ministerial Advisory Group (TEMAG) to "provide advice on how initial teacher education (ITE) programs could be improved to better prepare new teachers with the practical skills needed for the classroom" (Craven et al., 2014, p. 1) is another. Acceptance of the TEMAG report on ways to strengthen ITE, through stronger national oversight of standards and procedures for ITE programs (AITSL, 2015a; revised 2018, 2019) in collaboration with state and territory teacher regulatory authorities, and the call for, and implementation of, a teaching performance assessment (TPA) to demonstrate that graduating ITE students were 'classroom ready' is another.

DOI: 10.4324/9780429318504-6

However, as in other countries, including the United States, the introduction of such a high-stakes assessment raises issues for a variety of stakeholders: preservice teachers, teacher educators, regulatory and education authorities, and school communities including supervising teachers, classroom students, and parents. Referring to the United States, the goals of TPAs are both to improve individual teachers and to enhance teacher preparation (Darling-Hammond et al., 2013; Feuer et al., 2013).

The implementation of such assessments for licensure in the United States has led to numerous controversies that have raised several issues relating to the validity of the assessment, equity, and reliability for the purposes of individual credentialing and for accountability for teacher education institutions. Much of this debate has centered on the Educative Teacher Performance Assessment (edTPA; American Association of Colleges for Teacher Education [AACTE], 2017b; Stanford Center, 2018), reported to be used in 900 ITE programs across 41 states (De Voto et al., 2020; Gitomer et al., 2019; Cochran-Smith et al., 2013; Reagan et al., 2016).

This chapter examines the design and implementation of an Australian TPA, the Graduate Teacher Performance Assessment (GTPA®),[1] and other potential TPAs, through a broad set of principles concerning 'fairness'. It utilizes evidence from the U.S. context to illuminate issues of challenge and opportunity for Australia in implementing its teaching performance initiatives. The chapter lays out the broad sets of principles and requirements concerning fairness and effectiveness and then applies these to consider the design and implementation of the GTPA. It addresses social science issues for the implementation of Australia's policy for a TPA.

Fairness principles

Expectations for Australian TPAs are that they be "valid, reliable and moderated" (AITSL, 2015a, p. 10). However, the introduction of a new form of professional licensure can introduce greater social expectations about the quality of the licensure process and its impact on individuals and society. We therefore take a broader perspective to examine the GTPA in terms of 'fairness'. Fairness in assessment has been addressed in the United States in both the professional literature (Dorans & Cook, 2016) and in professional technical standards developed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (AERA/APA/NCME, 2014). Fairness is not a technical term and can mean different things to different people and in different contexts (AERA/APA/NCME, 2014). While it encompasses technical aspects of assessment and test design, it can involve notions of fair treatment embedded in the law, but it can also mean just treatment in a broad sense and is perhaps most often recognized when people perceive unfairness in the treatment of themselves or others (Dorans & Cook, 2016).

In the United States, the AERA, the APA, and the NCME have jointly developed and published technical standards for assessment development and implementation (hereafter, *Test Standards*) (AERA/APA/NCME, 1999, 2014). In the absence of similar standards in other countries such as Australia, it is common for testing professionals to refer to the U.S. standards for guiding principles.[2]

The 2014 *Test Standards* set out provisions for ensuring fairness, but these are closely interwoven with broader technical considerations of the quality of the instrument and of the context in which assessment occurs. These professional technical standards apply to both tests and assessments and to their uses to make judgments about individual performance (or education programs). There are cornerstone technical concepts concerning such issues as validity and reliability and provisions specific to fairness issues. The technical standards for testing and assessment specify that:

> a test that is fair within the meaning of the *Standards* reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended or unintended construct.
>
> *(AERA/APA/NCME, 2014, p. 50)*

Fairness is an issue applicable primarily to future teachers participating in a TPA but fairness issues also apply to ITE programs and personnel. This chapter additionally highlights fairness as a consideration for the public at large, which is entitled to fair outcomes from an education reform initiative. As a result, fairness will be addressed here in terms of the quality of the TPA as an assessment technology. Fairness will also be addressed in terms of the learning opportunities provided to preservice teachers to prepare them to take a TPA as well as an opportunity to challenge an outcome they consider unfair. The learning opportunities issue will also be addressed from the perspective of ITE programs and faculties and from the perspective of cooperating teachers who oversee field placements of future teachers. Fairness will also be assessed in terms of the processes and procedures in place to address potential errors in decision-making or injustice in the outcomes of decisions based on the assessment.

Fairness by ensuring the technical quality of the assessment

It is not uncommon for many nations in the current era to use assessments as a primary policy tool for driving education reform. Such an approach, of course, places great weight on the quality of the assessment utilized and the fidelity of the system in implementing the assessment. The technical quality of an assessment and the defensibility of the inferences drawn from assessment results are critical aspects of fairness. Technical quality here relates to two uses of a TPA, first, as an individual assessment for future teachers for graduation and credentials to teach and, second, as a type of "soft" governmental requirement for accreditation of university institutions' ITE programs.

The *Test Standards* include an entire chapter on fairness (AERA/APA/NCME, 2014, chapter 12). Fairness is identified as a "fundamental issue in protecting test takers and test users in all aspects of testing" (p. 49). The *Standards* set forth provisions on fairness that address: fair treatment of test-takers during the testing process; fair treatment as assured by the quality of the measurement; fairness as the absence of measurement bias; fairness as access to the construct measured; and fairness as validity of score interpretations and use (p. 51). Fairness is critical to the validity of the inferences drawn from an assessment.

The *Test Standards* also articulate a set of considerations and standards specific to educational assessment (chapter 12; AERA/APA/NCME, 2014). These are applicable to a TPA and relate to the design and development of educational assessments; use and interpretation of educational assessments; and the administration, scoring, and reporting of educational assessments. The *Test Standards* include a chapter (chapter 13) addressing program evaluations and another on credentialing (chapter 11). All of these chapters are applicable to a TPA and its use for decisions concerning future teachers and also the accreditation of ITE programs.

The Fairness Standards call for consideration of the various ways scores are reported, used, and interpreted, as well as consideration of the governance and legal system in which testing is practiced (AERA/APA/NCME, 2014, p. 49). Fairness for all test takers is regarded as a fundamental validity issue, with considerations for people with disabilities and those with diverse linguistic and cultural backgrounds (p. 49).

At the foundation of all professional technical standards are provisions for validity and reliability; both are essential to the fairness and effectiveness of an initiative.

Validity standards

The first cornerstone for test/assessment quality is validity. The *Test Standards* describe validity as

> …the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence for each interpretation is needed.
>
> *(AERA/APA/NCME, 2014, p. 225; see also pp. 11–31, 195, 210–211)*

A clear articulation of the evidentiary argument for substantiating the inferences drawn from a performance assessment is required (Lane & DePascale, 2016). For Australia's TPAs, the validity argument focuses upon the APST (AITSL, 2011), the *Accreditation of Initial Teacher Education Programs in Australia: Standards and Procedures* (AITSL, 2015a), and classroom readiness of ITE graduates (see Chapter 3 within this book). These policy requirements are designed to achieve consistency of outcomes across accredited ITE programs.

Validity considerations for a TPA must address the constructs of teacher readiness and ITE program quality. Content and construct validity concerns classroom readiness and representation of the nation's professional standards for graduating teachers and accreditation requirements for ITE programs. Because performance on the TPA is high-stakes, consequential validity evidence is important. Finally, evidence of the extent to which the TPA has predictive validity will become important.

The relationship between validity and reliability considerations must be addressed as well. So, for example, in the U.S. context, when the reliability of one TPA, the Performance Assessment for California Teachers (PACT) was questioned, researchers called for further consequential validity evidence for PACT, given the reports of limited inter-rater reliability of that assessment (Porter & Jelinek, 2011).

The ongoing research on the validity of TPAs in the United States has raised many conclusions, findings, and calls for the need for further research (Goldhaber et al., 2017). For example, while a correlation between academic course grades and methods course grades and PACT performance can be found (Sandholtz & Shea, 2015), there may not be a correlation between grades in student teaching and university supervisors' predictive ratings for future teachers and scores on a performance assessment (PACT), particularly at the highest and lowest ends of the performance scale (Gitomer et al., 2019; Sandholtz & Shea, 2012, 2015).

Reliability standards

The second cornerstone for test/assessment quality is reliability. The *Test Standards* describe reliability as consistency in measurement:

> …the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group.
>
> *(AERA/APA/NCME, 2014, pp. 222–223; see also pp. 33–47)*

There are some U.S. experiences with efforts to address reliability in a TPA. California's PACT uses an approach to inter-rater reliability that requires institutions to use double scoring of a portion of each institution's group of submissions; there was, in one study, considerable variation among scorers in the pass/failing distinction, especially in the failing designation raising substantial concerns about reliability (Gitomer et al., 2019; Porter & Jelinek, 2011). PACT has used a calibration training exercise in which scorers must meet a required standard before scoring (Sandholtz & Shea, 2015).

Standard setting and scoring

The *Test Standards* include a chapter addressing workplace testing and credentialing (chapter 11), which includes considerations for setting standards for licensure or certification testing (AERA/APA/NCME, 2014, pp. 169–177, pp. 95–110).

> Defining the minimum level of knowledge and skill required for licensure or certification is one of the most important and difficult tasks facing those

> responsible for credentialing. The validity of the interpretation of the test scores depends on whether the standard for passing makes an appropriate distinction between adequate and inadequate performance.
>
> *(p. 176)*

Two critical factors in standard setting are *who* decides the cutoffs for scoring and *how* they decide.[3] In the United States, for the edTPA, once scored, the passing standard is set by individual states that use edTPA scores. As a result, passing scores can be extremely variable from state to state (American Association of Colleges for Teacher Education, 2017b; Sawchuk, 2015). The processes used in Australia for the GTPA, as outlined in Chapters 6 and 7 within this book, focus on consistency in the application of the established standard across institutions and the resolution of identified differences.

In the past several years in the United States, there has been an ongoing concern among some teacher educators about the edTPA, particularly once scoring was given to Pearson[4]. These concerns have focused on contests over the scoring of the candidates' submissions, but there have also been broader assertions arising from the perception that the edTPA intrudes on the judgment of university faculty and undercuts the value of the professional contributions and judgments of cooperating practitioners in the schools (Cochran-Smith et al., 2016; Greenblatt, 2017; Greenblatt & O'Hara, 2015; Jordan & Hawley, 2016). Another concern has been Pearson's reduction in the number of scorers per TPA to one (Gitomer et al., 2019; Berlak & Madeloni, 2015).

In the United States, part of the source of the controversy is a strong cultural tradition of academic freedom for university faculty to determine their own curriculum, pedagogy, and evaluation of students (Pullin, 2004). In Australia, through the academic institution-based approach to the GTPA, supported by internal moderation and cross-institutional moderation online (CIM-Online™),[5] group meetings, and safeguards of standards, the GTPA is designed to ensure that professional responsibility for judgments lies with the teacher educators from each participating university. They undertake these activities in the national group known as the Collective.

Treatment of special populations

The impacts of the GTPA on both special populations of future teachers as well as the populations of special students the teachers serve, are another important set of issues and a particular design focus of TPAs. These relate directly to technical considerations of validity, reliability, standard setting, assessment administration, and implementation of a TPA. The *Test Standards* offer a chapter on fairness issues involving special populations of test-takers (AERA/APA/NCME, 2014, chapter 4, pp. 49–74). The impact of a TPA on special populations needs to be considered in the context in which it is used, as well as in terms of its promotion of opportunities to learn for future teachers and for the students they serve, as will be discussed in the next section of this chapter.

Given the situatedness of the GTPA within universities, the extent to which adjustments or accommodations are made available to ITE students will be affected by the provisions and policies of individual universities. As will be discussed in Chapter 10 within this book, the GTPA in Australia and university provision are governed by anti-discrimination legislation, specifically the *Disability Discrimination Act 1992* (Cth), and accompanying subsidiary legislation, the *Disability Standards for Education 2005* (Cth) [DSE]. The latter establish clear expectations that all students should be able to access education provision, and demonstrate their learning, without prejudice. This incorporates the provision of appropriate adjustments in assessment (DSE, s 3.4). However, as will be discussed in Chapter 10 within this book, the DSE also note that adjustments have to be 'reasonable' in terms of financial hardship and impact on others, and the 'integrity' of an assessment for certification purposes should be maintained (s 3.4.3 *Note*). A further issue related to special populations is whether there has been any adverse impact on specific student groups such as students from socio-economic disadvantaged backgrounds, and those with language backgrounds other than English. Drawing on the experience of the United States, nonpublication of institutional reports will offset potential misuse of GTPA outcomes to establish 'league tables' of institutions or other inappropriate or unintended use of the data (AERA/APA/NCME, 2014, p. 18).

One critique of TPAs in the United States is that current initiatives to reform ITE, including edTPA, take a "thin equity" approach (Cochran-Smith et al., 2016). Cochran-Smith et al. (2016) concluded that a "strong equity" approach is needed to address the needs of diverse learners in a meaningful way. This entails "focusing directly on creating the conditions for high-quality teaching" and "include[s] preparing and expecting teachers to: recognize and build on the knowledge traditions of marginalized groups; understand and challenge inequities in the existing structures of schools and schooling; and work with others in larger efforts for social justice and social change." (p. 4). As will be discussed in Chapters 7 and 8 within this book, the wide participation of teacher educators from many different institutions in the design, implementation, and scoring of the GTPA is intended to have a salutary impact on enhancing teacher education programs. Further, a discussion of the evidence of a washback effect on curriculum in teacher education programs and preparation, which lead to opportunities to learn, is presented in Chapter 9.

Opportunity to learn

Ideally, an assessment can promote learning opportunities (Darling-Hammond et al., 2013; Gearhart & Osmundson, 2009; Moss et al., 2008). An important consideration for any high-stakes performance assessment is whether students have been provided an adequate opportunity to learn what is covered on the assessment (Lane & DePascale, 2016; Linn, 1994). Important technical considerations for the design and implementation of a TPA relate to fairness in terms of ensuring a meaningful opportunity to learn for future teachers. According to the *Test Standards* (2014), "opportunity to learn—the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test" is an important component of fairness in testing (AERA/APA/NCME, 2014, p. 56).

> … it is generally accepted that before high-stakes consequences can be imposed for failing an examination in educational settings, there must be evidence that students have been provided curriculum and instruction that incorporates the constructs addressed by the test.
>
> *(p. 57)*

The formal name of the edTPA, the Educative Teacher Performance Assessment, implies a formative function for the assessment that has been widely adopted in the United States. In California, PACT has been reported to play a key role for future teachers and teacher educators in defining learning opportunities (Reagan et al., 2016; Wei & Pecheone, 2010). To the extent a TPA reflects new content standards for future teachers and for teacher preparation programs, sufficient chance is required to teach to the new standards. This provision is critical, both in terms of full understanding of the standards, resources, and capability to present curriculum and instruction for the standards, and time to prepare (Pullin, 2001, 2015). Some TPA implementation, including the edTPA, may be in part an appropriate reflection of the insufficiency of learning opportunities to prepare for the assessment. These could involve low passing scores, low passing rates, and changing score requirements (DeVoto et al., 2020).

As part of the implementation of the GTPA, universities have used data generated from the scoring and CIM-Online™ processes to identify areas of programs that need strengthening to ensure students have been prepared to undertake their assessment. This has therefore led to a 'washback' effect on curriculum in ITE programs and preparation, increasing the validity of the GTPA and programs in terms of the AITSL requirements for a TPA, as discussed in Chapters 4 and 9 within this book. This has enhanced ITE students' opportunity to learn within university programs and to be certified as 'classroom ready' (McDowall et al., 2021).

However, responsibilities for ensuring adequate and meaningful opportunity to learn what a teacher needs to know and be able to do, as measured by a TPA, rest with several different actors beyond universities' responsibility for their own curriculum and requirements. These include the schools which provide placement sites for student fieldwork, and jurisdictions where universities have limited control over opportunities for students. AITSL identifies "key groups that share responsibility for the professional experience component of initial teacher education programs, which include: professional experience sites, supervising teachers, pre-service teachers, providers of initial teacher education, education systems" (AITSL, 2015b, p. 2).

Guidelines for the GTPA indicate responsibilities for key stakeholders in terms of the opportunity to learn, emphasizing collaboration and communication. Universities are responsible for liaising with schools to ensure all participants are informed about the purposes and timing of the assessment. Advice is given that "Roles and responsibilities need to be addressed to assure schools that the GTPA is not an additional workforce demand, but rather a part of the further strengthening of ITE programs" (Institute for Learning Sciences and Teacher Education [ILSTE], 2020a, p. 4). Of industrial interest is that teacher educators are responsible for the assessment and moderation of the GTPA (ILSTE, 2020b). This highlights that the role of the supervising teacher (school-based mentor)

> is primarily consultative and advisory in nature… for example, assisting the preservice teacher to select student data representing the range of capabilities in the class and the relevant achievement standards …[and] undertaking moderation discussions to review the use of standards and criteria in assessing student work.
>
> *(ILSTE, 2020b, p. 3)*

AITSL, registration and education authorities, ILSTE as the host of GTPA, and individual universities have provided online resources to guide supervising teachers in student placement supervision, recording of evidence, and the implementation of TPAs (e.g., AITSL, 2017b; ILSTE, 2020c; Queensland College of Teachers, n.d.-a, n.d.-b).

Other fairness considerations

The 2014 *Test Standards* note that fairness principles, especially for high-stakes decision-making, include multiple opportunities for students to demonstrate their knowledge, alternate forms, and use of "multiple criteria rather than just a single test score" (AERA/APA/NCME, 2014, p. 186). Although the GTPA results in a single judgment as to whether a student meets the necessary standard, this judgment is derived from a complex performance task that integrates evidence from a range of sources. The rubric for assessing performance draws on multiple criteria reflecting AITSL expectations. When there are multiple uses for a test or assessment (such as for both individual and institutional data) then validity, reliability, and fairness evidence is needed for each use (pp. 188, 195) and when used for a purpose not intended by the developer, that user must provide the necessary evidence (p. 195). Universities may score and use GTPA outcomes to contribute to overall grade point average (GPA) calculations. However, the primary purpose of the GTPA, and work on validation and reliability that have been undertaken, is the designation that the graduating teacher is classroom ready.

Different institutions have in place different mechanisms to assist students who are not immediately successful. These may include revision of core components of the GTPA assessment task, as well as opportunities to repeat the GTPA in full. The number of times a student may repeat the GTPA within a program will be governed by the university. However, the student will not be eligible for graduation and teacher registration until achievement on the GTPA is satisfactory.

Assessing fairness of the GTPA

The technical issues presented above suggest a series of considerations about GTPA design and implementation. Given the technical and social science issues related to TPAs and their implementation, and the experiences with TPAs in the U.S. context, it is pertinent to consider the issues that may arise for the GTPA and TPAs more generally in the Australian context. Fairness can be seen as adherence to the *Test Standards*, including both the explicit fairness provisions and the broad range of technical requirements. Fairness in testing requires an assessment of high technical quality. Consideration of fairness and the GTPA that follows here will focus on the use of the GTPA for graduation from ITE programs, but issues will also arise to the extent the GTPA may be utilized for government accreditation of university ITE programs.

Applying the fairness principles to the GTPA

In its call for the introduction of a TPA, Australia incorporates requirements for the validity and reliability of assessments for uses for future teacher graduation determinations and for program accreditation. The *Accreditation of Initial Teacher Education Programs in Australia: Standards and Procedures* (AITSL, 2015a) (National Program Standards) requires that a preservice teacher satisfactorily completes a TPA in their final year to graduate (see Program Standard 1.2). For accreditation purposes, the TPA, as noted, must be "valid, reliable and moderated" (AITSL, 2015a, p. 10). Guidelines on elements and processes that have to be met for a TPA to receive AITSL endorsement have been published (AITSL, 2017a). The GTPA, including the established standard, was assessed against Program Standard 1.2 (AITSL, 2015a) by an AITSL-led expert panel in 2017–2018. As an outcome of the expert panel assessment, the GTPA was officially endorsed as meeting the established requirements of Program Standard 1.2, subject to program-level consideration which should be evidenced by individual providers who intend to use the GTPA. These conditions pertain to the fidelity of implementation and evidence (ILSTE, 2020c).

Effectiveness principles

The technical quality of an assessment and the defensibility of the inferences drawn from assessment results are critical aspects of fairness, as discussed above, but also are essential to the effectiveness of an initiative. Failure to embrace, support, fund, and fully implement an effective education reform can be seen as a denial of fairness to those who participate in implementing the system or who were the intended beneficiaries of the system. Introducing a large-scale and expensive education reform is not worth the effort unless there is reasonable expectation that the reform will improve educational opportunities. If a performance assessment system does not adequately distinguish between acceptable performance and unacceptable performance of future teachers or ITE programs, the TPA will not be an effective policy tool for reform.

What makes a reform effective? In the most recent international comparative study of ITE, the conclusion was reached that no single innovation can improve teacher quality; teacher innovations and the way they fit together in the entire system of education in a country or jurisdiction are key to successful improvement of teacher quality. This "teaching and learning system" and the coherent and well-fit pieces of the entire complex system of both policies and policy implementation are required for meaningful reform of the teaching profession (Darling-Hammond et al., 2017; Pullin, 2001, 2014a, 2015). Scholars have also concluded in studies of other areas of education that effective reform requires a continuous improvement approach (Bryk et al., 2016). What will the outcome be for Australian reforms and the introduction of the TPA? The overall purpose of development of stronger ITE program accreditation principles, including assessment of graduate performance, is to ensure "quality", "classroom ready" graduates who will have a "positive impact on student learning from day one in their first teaching role" (AITSL, 2017b, n.p.).

Assessing effectiveness of the GTPA

There is no reason to implement a reform policy unless the tools utilized to enact the reform are of sufficient quality to be effective. In assessing the effectiveness of a competence assessment like the GTPA, of primary importance are the stated goals of making graduate teachers more classroom ready and ITE programs more consistent and accountable. Will the GTPA ultimately further the stated goals of the initiative?

To what extent does an endorsed TPA have predictive validity, that is, to what extent does it reflect eventual performance on the job as a teacher? There is some evidence on performance assessment in the United States that some assessment does not in fact predict future professional performance (see, for example, Grissom et al., 2017, concerning the limited evidence of predictive validity for an assessment of school principals).

Implementation of TPAs in Australia is recent; little evidence is available about their impact on the quality of teaching and learning in schools. It is therefore premature to assert TPA effectiveness as a policy lever to improve teaching effectiveness, and in turn, learning outcomes. To this end, the GTPA includes design features for a continuous improvement approach; universities use data for curriculum review and program renewal. This includes data generated through rigorous, large-scale cross-institutional online scoring and moderation.

Effectiveness might also usefully be assessed from the perspective of a broader set of public benefits. In the U.S. context, teacher quality indicators have also been used as indicators of effectiveness for determining the success of an entire university program, not just its ITE components. Teacher quality indicators have also been used as evidence for resolving some state constitutional disputes over whether schools are providing fair funding and resources by state legislatures (Pullin, 2001, 2014a, 2015).

Fairness and effectiveness are considerations in social science, as well as reflections of the public's notions and policy-makers' choices of what is for the common good in education. Fairness and effectiveness, or more appropriately, perceptions of lack of fairness or failures to achieve effectiveness, can also lead to legal disputes, discussed from the perspective of U.S. experience and potential dispute processes in legal jurisdictions in Australia, as described in Chapter 10.

Conclusion

In the United States, there have been decades of efforts to improve ITE; wave after wave of reforms have been initiated, yet persistent concerns about teacher education and teacher quality remain (Pullin, 2017). In many respects, the current Australian approach, through initiatives like the GTPA, present a promising prospect. The GTPA is one component of an effort to take a systemic approach to achieve meaningful reform of the teaching profession and ITE. Is it sufficient to effectively enhance the opportunity to learn and the goals of the reforms? Will it do so both fairly and effectively? Legal controversies in the United States have often resulted from these types of initiatives, as will be discussed in Chapter 10. For Australia, will recent efforts to reform ITE and entry to the teaching profession provoke similar controversies? To date, no legal controversies have arisen in Australia in response to the introduction of the new competence assessment requirement. However, overall, the uptake of TPAs across the country has been slow. Significant issues remain to be addressed, including methodologies for national benchmarking and the feasibility of establishing an agreed national standard (see Chapter 11 within this book).

The answers to these questions will depend, in large part, on the choices that have been made and continue to be made about the design and implementation of TPAs and the other aspects of the current Australian education reform initiatives. Both United States and international scholars have noted the importance of a systemic approach (Darling-Hammond et al., 2017) and an "improvement science" approach (Bryk et al., 2016) to education reform. The success of the GTPA as well as the other components of current reform in Australia will depend in large part on the extent to which government, scholars, and practitioners are determined to stay the course and commit the resources for the difficult work to be done. Included in this work must be efforts to implement ongoing discernment of potential legal challenges to these initiatives, as discussed in Chapter 10 within this book, and how, or whether, to respond to those.

Notes

1 Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

2 Australian reference to principles and attributes of good assessment focus more generically on teacher classroom assessment (see, for example, Queensland Curriculum and Assessment Authority, 2018) which identifies general principles of alignment of assessment with curriculum, pedagogy and reporting, equitable assessment for all students,

evidence-based judgment, and decision-making through "a range and balance of tasks over time", continuous over time, transparency, and informative. The preparation and analysis of data from Australia's national literacy and numeracy assessments, NAPLAN, refer to test development quality more obliquely, for example, noted as "comprehensive, rigorous and draw[ing] on the best available expertise within Australia and national and international best practice" including "processes associated with the Programme for International Student Assessment (PISA)" (Australian Curriculum, Assessment and Reporting Authority, 2013, pp. 1–2).

3 Some information in relation to the GTPA and standard setting, scoring and assessment fidelity is available in papers by Wyatt-Smith et al. (2020, 2021) and Adie and Wyatt-Smith (2020) and in Chapters 4, 6 and 7 within this book.

4 Pearson is a global corporation that provides learning and assessment resources and services.

5 Acknowledgment: The online model of cross-institutional moderation (CIM-Online™) was conceptualized and developed in the Institute for Learning Sciences and Teacher Education, Australian Catholic University. For a discussion of CIM-Online™, readers are advised to also see Wyatt-Smith et al. (2021).

References

Adie, L., & Wyatt-Smith, C. (2020). Fidelity of summative performance assessment in initial teacher education: The intersection of standardisation and authenticity. *Asia-Pacific Journal of Teacher Education*, *48*(3), 267–286. https://doi.org/10.1080/1359866X.2019.1606892

Akiba, M. (2017). Understanding cross-national differences in globalized teacher reforms. *Educational Researcher*, *46*(4), 153–168. https://doi.org/10.3102/0013189X17711908

American Association of Colleges for Teacher Education (AACTE). (2017a, November 3). News release: Annual administrative report on EdTPA data show continued growth and support. edtpa.aacte.org/news

American Association of Colleges for Teacher Education (AACTE). (2017b). *State edTPA Policy Overview*. http://edtpa.aacte.org/state-policy.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Australian Curriculum, Assessment and Reporting Authority. (2013). *NAPLAN technical summary*.http://webcache.googleusercontent.com/search?q=cache:kW3m7MBtCsUJ:www.aph.gov.au/DocumentStore.ashx%3Fid%3D802bd846-eb27-40ca-a515-8611d2f6d172+&cd=2&hl=en&ct=clnk&gl=au&client=safari

Australian Government Department of Education and Training. (2016). *Quality schools, quality outcomes*. https://docs.education.gov.au/node/40671

Australian Institute for Teaching and School Leadership (AITSL). (2015a; revised 2018, 2019). *Accreditation of initial teacher education programs in Australia: Standards and procedures*. https://www.aitsl.edu.au/docs/default-source/national-policy-framework/accreditation-of-initial-teacher-education-programs-in-australia.pdf?sfvrsn=e87cff3c_28

Australian Institute for Teaching and School Leadership (AITSL). (2015b). *Professional experience. Participant roles and responsibilities*. https://www.aitsl.edu.au/docs/default-source/default-document-library/professional-experience---participant-roles-and-responsibilities19e58791b1e86477b58fff00006709da.pdf?sfvrsn=e9c3f53c_0

Australian Institute for Teaching and School Leadership (AITSL). (2017a). *Teaching performance assessment*. https://www.aitsl.edu.au/deliver-ite-programs/teaching-performance-assessment

Australian Institute for Teaching and School Leadership (AITSL). (2017b). *Teaching performance assessments: An overview for schools*. https://www.aitsl.edu.au/deliver-ite-programs/teaching-performance-assessment/teaching-performance-assessments-an-overview-for-schools

Australian Institute for Teaching and School Leadership (AITSL). (2011; revised 2018). *Australian Professional Standards for Teachers*. https://www.aitsl.edu.au/teach/standards

Berlak, A., & Madeloni, B. (2015). From PACT to Pearson: Teacher performance assessment and the corporatization of teacher education. In P. R. Carr & B. J. Porfilio (Eds.), *The phenomenon of Obama and the agenda for education: Can hope audaciously trump neoliberalism*? (pp. 193–214). IAP Information Age Publishing.

Bryk, A., Gomez, L., Grunow, A., & LeMahieu, P. (2016). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.

Burns, D., & McIntyre, A. (2017). *Empowered educators in Australia: How high-performing systems shape teaching quality*. Jossey Bass.

Cochran-Smith, M., Piazza, P., & Power, C. (2013). The politics of accountability: Assessing teacher education in the United States. *The Educational Forum*, 77(1), 6–27. https://doi.org/10.1080/00131725.2013.739015

Cochran-Smith, M., Stern, R., Sánchez, J. G., Miller, A., Keefe, E. S., Fernández, M. B., Chang, W., Carney, M. C., Burton, S., & Baker, M. (2016). *Holding teacher preparation accountable: A review of claims and evidence*. National Education Policy Center. http://nepc.colorado.edu/publication/teacher-prep

Craven, G., Beswick, K., Fleming, J., Fletcher, T., Green, M., Jensen, B., Leinonen, E., & Rickards, F. (2014). *Action now: Classroom ready teachers – Report of the Teacher Education Ministerial Advisory Group (TEMAG)*. Australian Government. https://docs.education.gov.au/documents/action-now-classroom-ready-teachers-report.

Darling-Hammond, L., Burns, D., Campbell, C., Goodwin, A. L., Hammerness, K., Low, E., McIntyre, A., Sato, M., & Zeichner, K. (2017). *Empowered teachers: How high-performing systems shape teaching quality around the world*. Jossey-Bass.

Darling-Hammond, L., Newton, S., & Wei, R. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation and Accountability*, 25(3), 179–204. https://doi.org/10.1007/s11092-013-9163-0

De Voto, C., Olson, J., & Gottlieb, J. (2020). Examining diverse perspectives of edTPA policy implementation across states: The good, the bad, and the ugly. *Journal of Teacher Education*, *72*(1), 1–14. https://doi.org/10.1177/0022487120909390

Disability Discrimination Act 1992 (Cth).

Disability Standards for Education 2005 (Cth) [DSE].

Dorans, N. J., & Cook, L. L. (Eds.). (2016). *Fairness in educational assessment and measurement*. Routledge.

Feuer, M. J., Floden, R. E., Chudowsky, N., & Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options*. National Academy of Education. www.naeducation.org.

Gearhart, M., & Osmundson, E. (2009). Assessment portfolios as opportunities for teacher learning. *Educational Assessment*, *14*(1), 1–24. https://doi.org/10.1080/10627190902816108

Gitomer, D. H., Martínez, J. F., Battey, D., & Hyland, N. E. (2019). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal*, (*58*)1, 3–31. https://doi.org/10.3102/0002831219890608

Goldhaber, D., Cowan, J., & Theobald, R. (2017). Evaluating prospective teachers: Testing the predictive validity of the edTPA. *Journal of Teacher Education*, *68*(4), 377–393. https://doi.org/10.1177/0022487117702582

Greenblatt, D. (2017, April). The contradictions and consequences of the edTPA. *Paper presented at the Annual Meeting of the American Educational Research Association*, San Antonio, Texas.

Greenblatt, D., & O'Hara, K. E. (2015). Buyer beware: Lessons learned from edTPA implementation in New York state. *Teacher Education Quarterly*, *42*(2), 57–67.

Grissom, J., Mitani, H., & Blissett, R. (2017). Principal licensure exams and future job performance: Evidence from the School Leaders Licensure Assessment. *Educational Evaluation & Policy Analysis*, *39*(2), 248–280. https://doi.org/10.3102/0162373716680293

Institute for Learning Sciences and Teacher Education [ILSTE]. (2020a). Fact sheet: Information for schools and supervising teachers. Australian Catholic University (ACU). https://www.graduatetpa.com/wp-content/protect/2021/Information%20for%20Schools%20and%20Supervising%20Teachers.pdf

Institute for Learning Sciences and Teacher Education [ILSTE]. (2020b). Fact sheet: Information for teacher educators and higher education institutions. Australian Catholic University (ACU). https://www.graduatetpa.com/wp-content/protect/2021/Information%20for%20Teacher%20Educators%20and%20HEIs.pdf

Institute for Learning Sciences and Teacher Education [ILSTE]. (2020c). Graduate teacher performance assessment. Australian Catholic University (ACU). https://www.graduatetpa.com/

Jordan, A. W., & Hawley, T. (2016). By the elite, for the vulnerable: The edTPA, academic oppression, and the battle to define good teaching. *Teachers College Record, Number:19461.*

Lane, S., & DePascale, C. (2016). Psychometric considerations for performance-based assessments and student learning objectives. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 77–106). Routledge.

Linn, R. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, *23*(9), 4–14.

McDowall, A., Mills, C., Cawte, K., & Miller, J. (2021). Data use as the heart of data literacy: An exploration of pre-service teachers' data literacy practices in a teaching performance assessment. *Asia-Pacific Journal of Teacher Education*, *49*(5), 487–502. https://doi.org/10.1080/1359866X.2020.1777529

Moss, P., Pullin, D., Gee, J., Haertel, E., & Young, L. (2008). *Assessment, equity and opportunity to learn*. Cambridge University Press.

Porter, J. M., & Jelinek, D. (2011). Evaluating inter-rater reliability of a national assessment model for teacher performance. *International Journal of Educational Policies*, *5*(2), 74–87.

Pullin, D. (2001). Key questions in implementing teacher testing and licensing. *Journal of Law and Education*, *30*(3), 383–429.

Pullin, D. (2004). Accountability, autonomy, and academic freedom in educator preparation programs. *Journal of Teacher Education*, *55*(4), 300–312. https://doi.org/10.1177/0022487104266717

Pullin, D. (2014a). *Performance, value and accountability: Public policy goals and legal implications of the use of performance assessments in the preparation and licensing of educators. Paper commissioned by the Council of Chief State School Officers and Stanford Center for Assessment, Learning, and Equity (SCALE).* https://secure.aacte.org/apps/rl/res_get.php?fid=1504&ref=edtpa

Pullin, D. (2014b). Professional test standards in the eyes of the law. *Educational Measurement: Issues and Practice*, *33*(4), 19–21.

Pullin, D. (2015). Performance measures for teachers and teacher education: Corporate education reform opens the door to new legal issues. *Education Policy Analysis Archives*, *23*(81), 1–32. https://doi.org/10.14507/epaa.v23.1980

Pullin, D. (2017). What counts? Who is counting? Teacher education improvement and accountability in a data-driven era. In J. Nuttall, A. Kostogriz, M. Jones, & J. Martin (Eds.), *Teacher education policy and practice: Evidence of impact, impact of evidence* (pp. 3–16). Springer.

Queensland College of Teachers. (n.d.-a). *Supervising professional experience*. https://www.qct.edu.au/teaching-in-queensland/supervising-professional-experience

Queensland College of Teachers. (n.d.-b). *Teaching performance assessments*. https://www.qct.edu.au/teaching-in-queensland/teaching-performance-assessment

Queensland Curriculum and Assessment Authority. (2018). *Principles of quality assessment*. https://www.qcaa.qld.edu.au/k-12-policies/student-assessment/understanding-assessment/principles-quality-assessment

Reagan, E. M., Schram, T., McCurdy, K., Chang, T.-H., & Evans, C. M. (2016). Politics of policy: Assessing the implementation, impact, and evolution of the Performance Assessment for California Teachers (PACT) and edTPA. *Education Policy Analysis Archives*, *24*(9), 1–23. https://doi.org/10.14507/epaa.24.2176

Sandholtz, J. H., & Shea, L. M. (2012). Predicting performance: A comparison of university supervisors' predictions and teacher candidates' scores on a teaching performance assessment. *Journal of Teacher Education*, *63*(1), 39–50. https://doi.org/10.1177/0022487111421175

Sandholtz, J. H., & Shea, L. M. (2015). Examining the extremes: High and low performance on a teaching performance assessment for licensure. *Teacher Education Quarterly*, *42*(2), 17–42.

Sawchuk, S. (2015, October 19). 18,000 would-be teachers took the edTPA last year: How did they do? *Education Week Blog*. http://blogs.edweek.org/edweek/teacher-beat/2015/10/2014_scores_on_edtpa.html

Stanford Center for Assessment, Learning and Equity (SCALE) (n.d.) *Frequently asked questions*. https://scale.stanford.edu/teaching/pact/faq

Stanford Center for Assessment, Learning and Equity (SCALE), the American Association of Colleges for Teacher Education (AACTE) and Evaluation Systems of Pearson (2018, September). *Educative assessment and meaningful support: 2017 edTPA administrative report*. https://secure.aacte.org/apps/rl/res_get.php?fid=4271&ref=edtpa

Wei, R., & Pecheone, R. (2010). Assessment for learning in preservice teacher education: Performance-based assessments. In M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality: A handbook* (pp. 69–132). Jossey-Bass.

Wyatt-Smith, C., Adie, L., & Nuttall, J. (Eds.). (2021). *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration*. Springer.

Wyatt-Smith, C., Humphry, S., Adie, L., & Colbert, P. (2020). The application of pairwise comparisons to form scaled exemplars as a basis for setting and exemplifying standards in teacher education. *Assessment in Education: Principles, Policy & Practice*, *27*(1), 65–86. https://doi.org/10.1080/0969594X.2020.171232

PART 2

Data analytics, systems thinking, and digital architecture

6

VALIDATION, RELIABILITY, AND STANDARD SETTING IN THE GTPA

A focus on methodologies and judgment

Introduction

This chapter describes the approach taken to explore the processes and practices relied upon to set the standard for graduate readiness to enter the teaching workforce. In writing the chapter, our primary aim was to share with readers the steps involved in enabling the teaching profession to combine their evaluative expertise for the purpose of achieving a shared understanding of what it means for a preservice teacher to be ready to teach in the classroom. Reaching agreement on the standard for preservice teacher competence to teach, by teacher educators from twelve universities who all brought different experiences from a range of contexts, was never going to be straightforward or achievable in a single session. It was necessary to identify strategies that would allow in-depth discussion and convergence of individual recommendations based on group evidence. The processes for harnessing the expertise of many participants required the iterative application of human judgment, data analysis, interpretation, and decision-making over eight days of purposeful workshop sessions. A key ingredient of the workshops was the genuine exchange of thoughts and ideas which was only possible through the willingness of teacher educators to reflect on their judgment choices. The steps involved in our approach are potentially transferable to standard setting in a range of professions.

The standard for readiness to teach had not been formulated at the time the Teacher Education Ministerial Advisory Group (TEMAG) review (Craven et al., 2014) called for a more intensive focus on the rigorous assessment of the quality of graduates from the more than 45 providers of initial teacher education (ITE) across the country (see Chapters 1 and 4). The trial of the Graduate Teacher Performance Assessment (GTPA®)[1] in 2017 was undertaken as a first step to setting the standard. Three questions to be answered in the trial were: *What is involved in establishing the standard? Can we reach an agreement on the standard? How do we make the standard visible*

DOI: 10.4324/9780429318504-8

to those in the profession? Finding answers to these questions was necessary to address the policy requirement for establishing a validated teaching performance assessment (TPA) with a standard for preservice teacher competence that was accepted by the teaching profession (Australian Institute for Teaching and School Leadership [AITSL], 2018). Data demonstrating that the instrument was valid and could produce reliable judgments were necessary for the instrument to be endorsed and able to be implemented nationally.

Deriving the standard and identifying GTPA performances that represented the minimum acceptable level for meeting the standard was a challenge that required examination and judgment of a large corpus of completed GTPA performance samples from a wide range of contexts. We brought together a group of some 80 experienced teacher educators to act as a guild of assessors (Sadler, 1989). They were well placed to draw on their knowledge and experience of making sound qualitative judgments and contribute to the validation of the standard by examining how well their judgments agreed with others in the gathered guild. The procedural question was how best to draw on this expertise to realize the goal of establishing and articulating the standard.

Arriving at an agreed standard by the teaching profession was inevitably going to take time. As mentioned above, there was no pre-existing gold-standard method for setting the performance standard for teacher education, nor was there a previously published level or reference point that was agreed by experts as showing the minimum acceptable level in performance. This presented a new horizon that the teaching profession in Australia was being asked to navigate.

A methodological strategy was required to address the sequential objectives of (1) identifying the characteristics of performance to be recognizable as competent (ready) for classroom practice, (2) determining the minimum acceptable performance level for the agreed standard, and (3) identifying illustrative examples of performances at the threshold of meeting the established standard. The two chosen methods were the dominant profile judgment method (DPJM; Plake et al., 1997), and the pairwise comparison method for comparative judgment of performances (Thurstone, 1927). This chapter provides the inside story of how the two methods were applied in novel ways by the guild of teacher educators to first address objective (1), followed by objectives (2) and (3) together, through undertaking two consecutive studies denoted as Study 1 and Study 2, respectively. These two judgment methods had not previously been applied together in the field of teacher education to establish professional competence.

Senior policy staff in the Queensland College of Teachers (QCT)[2] had examined the GTPA for content validity, assessing it against the Australian Professional Standards for Teachers (APST; AITSL, 2011; see Chapter 4). The team of experts involved in further validation processes included members of AITSL, the GTPA Steering Group, the QCT Principals' Engagement Reference Group, and senior representatives from the Queensland Teachers Union and the Independent Education Union. The one year in-field trial in 2017 was designed to include the two methodologically distinct studies mentioned earlier, with a focus on validation, standard setting, and moderation to demonstrate reliability. How teacher educators applied the instrument and the scoring rubric across diverse university settings was examined.

Thirteen universities agreed to participate in the trial. Study 1 provided the opportunity for the teacher educators to come together, for the first time, to collectively (1) examine the de-identified authentic performance samples supplied by preservice teachers,[3] (2) apply the scoring rubric and performance level descriptors (PLDs), and (3) derive the standard of performance and the minimum level that was acceptable to the profession. Participants were asked to disengage from pre-established notions of their university grading systems with which they were most familiar (e.g., letter grades: A–E, and percentage bands for categories of performance such as high distinction, distinction, credit, pass). This was new ground in ITE and a theoretically sound judgment process was required for deriving the accepted standard from the collaborative expertise of the teacher educators, policy leaders, and employment authorities, representing diverse university and state contexts.

Consistent judgment of a complex performance assessment used to determine professional competence can be challenging. As discussed in Chapter 1, this was a new ask in the field of teacher education. Reliability of judgments can vary depending on the interrelationships between the site and system contexts in which the assessment occurs (see Chapter 3), as well as the evaluative experience of the judges in applying the performance standard and in recognizing when this has been achieved (Cooksey et al., 2007; Freebody & Wyatt-Smith, 2004). Additional complexity and challenges for judgment consistency are introduced when the assessment is scored by teacher educators with differential levels of expertise and experience, and with exposure to different assessment policy contexts. Ensuring that the GTPA was scored consistently against the established standard to demonstrate effective teaching practices in the classroom, not only within a university but also across many universities, required a judgment analysis framework that considers the perspectives of human judgment as well as the systematic features of the judgment process. The centerpiece of the framework was to be the professional standards and the established standard as the external referent for determining competence. Our starting proposition was that this framework should be accessible to all teacher educators with responsibility for assessing preservice teaching performances within their own institutions.

Introducing Study 1 and Study 2: A focus on methodologies

In Study 1, the approach taken to determine the provisional recommended level of performance for the GTPA during the trial was to apply a multi-stage judgmental process combining elements of the DPJM (Plake et al., 1997) and the analytic judgment method (Plake & Hambleton, 2001). These methods entailed the systematic integration of theory, judgment, and empirical evidence to support decisions about preservice teachers' performances relative to a defined performance level. While there were several potential methods that could be used to recommend passing standards for performance assessments (see Hambleton & Pitoniak, 2006; Plake et al., 1997), the selected methods needed to consider the unique features of the GTPA as a complex integrated task. The DPJM is most suited for complex performance assessments that involve multiple interlocking criteria with varying degrees of performance. Another feature of the DPJM is its capacity to incorporate multiple sources of information to contribute to the decision. Procedurally, the method iteratively incorporates discussion and empirical evidence into judgments while permitting experts to make their recommendations using compensatory, conjunctive, or a combination of these decision rules. Moderation was understood to be critical in the application of this multi-stage judgmental process.

The analytic judgment method informed the DPJM through the process of iterative scoring of GTPA samples at the criterion level, analysis of the scores, group discussions on the inconsistencies in scores, and the revision of scores to reach a consensus on sample performance. The analytic component of the study focused on the examination of consistency in judgment of the GTPA samples as an overall Meets (or Does Not Meet) and discrimination among the criteria in their impact on performance. The results from the analytic judgment method were used to provide examples of different performance profiles and decision rules that were then discussed during the application of the DPJM to arrive at a description of the minimum acceptable performance level for the standard for achieving an overall pass, taken to be Meets.

It has been proposed in previous research (Sadler, 2009; Smith, 1989) that "purposefully selected exemplars that provide concrete referents or illustrations of how expected characteristics of quality have been applied in judgment" (Wyatt-Smith et al., 2020, p. 1) could enhance the dependability of judgments (Harlen, 2005) against a written description of the standard. Exemplars have also been used to build teacher familiarity with standards and increase the dependability of their judgments in schooling systems. Precedents for the use of exemplars to support teachers in schools to improve the reliability of their judgments of students' classroom work for assessment include student work samples made available online by the NSW Government Education Standards Authority (NESA). These enable all teachers to consistently apply the standard for the common grade scale in their classrooms (see educationstandards.nsw.edu.au).

On the close of Study 1, an agreed description of a provisional GTPA passing standard had emerged but it remained to identify exemplars of performances at the threshold of meeting the standard. Study 2 was planned to achieve this objective. The method of pairwise comparison had been applied to identify exemplars of student performance for the common assessment grade scale in schools (see Bramley & Gill, 2010; Heldsinger & Humphry, 2010, 2013; Humphry et al., 2017; Humphry & McGrane, 2015) but it had not previously been applied to an authentic teaching performance assessment in ITE. The pairwise comparison method was chosen to follow the first study as it combined human judgment and statistical analysis to identify the 'line in the sand' for acceptable performance. More specifically, the method was a defensible approach to identify exemplars of GTPA performance at the threshold level of performance as well as performances assessed to be above and below Meets.

Study 2 was conceptualized in two parts and combined two adjacent comparative judgment methodologies that together provided a novel approach to standard setting (Thurstone, 1927). The first part of the study used a pairwise comparison method to identify GTPA exemplars that could be placed on a scale representing the standard at different levels of performance. Teacher educators were presented with carefully selected pairs of performance samples and for each pair were asked to make a binary decision on which sample performed better. The binary data collected from all pairwise decisions were analyzed, and the samples were ordered from highest to lowest on a scale. A subset of samples was selected to represent performances from across the range of the scale. The second part of the study considered judgments of the selected subset of performances where the transition between meeting and not meeting the standard was identified by an expert panel of experienced teacher educators (see Wyatt-Smith et al., 2020, for a detailed discussion on the application of the method). In this way, exemplars of the minimum acceptable performance level for the standard were established for reference by teacher educators judging GTPA performances.

The remainder of this chapter is written in three parts. The first part presents the design of Study 1, and the second part presents the design of Study 2. For each Study, the approach to analysis is described and the results presented. The third part discusses key insights into the application of the two methods to establish if and how they were complementary. At issue is the significance of the choice of the method: Would the result be different if only one method was applied to set the standard?

Processes and methods

Teacher educators from the participating universities contributed to the complete range of activities associated with the one-year GTPA trial including (1) implementation of the GTPA consistent with required conditions, (2) the submission of de-identified graded samples, (3) participation in monthly online meetings to support implementation and address emerging issues, and (4) completion of scoring activities during in-person meetings in Brisbane (6 days), scoring performances online using pairwise comparisons (approximately 1.5 days), and attendance at the Expert Group meeting in Brisbane (1 day). This time commitment was in addition to the initial scoring of GTPA performances within the host universities.

To undertake Study 1 and Study 2, it was first necessary to compile a large pool of completed GTPA performance samples representing a diverse range of contexts. Preservice teachers submitted their completed GTPA performances to their universities during the initial GTPA implementation in the first half of 2017, and teacher educators selected internally moderated and scored samples for contribution to cross-institutional judgment in the trial studies. The methodological design for selecting the samples was critical to ensure that one common performance standard was established across all possible teaching contexts. The samples were selected to be representative of student placement by the geographical location of the placement school, and the focus of the GTPA, including phase of schooling and teaching area. This required the careful management of sample provision and the processing of samples on receipt by the research team. For each of their teacher education programs, universities agreed to provide multiple performance samples with a mix of outcomes relative to their understanding of the standard including low level meets, high meets, high 'does not meet', and very low 'does not meet'. To support the universities in their provision of samples for the trial, guidelines were developed to inform sample selection and sample preparation for the electronic transfer of the samples to the trial database.

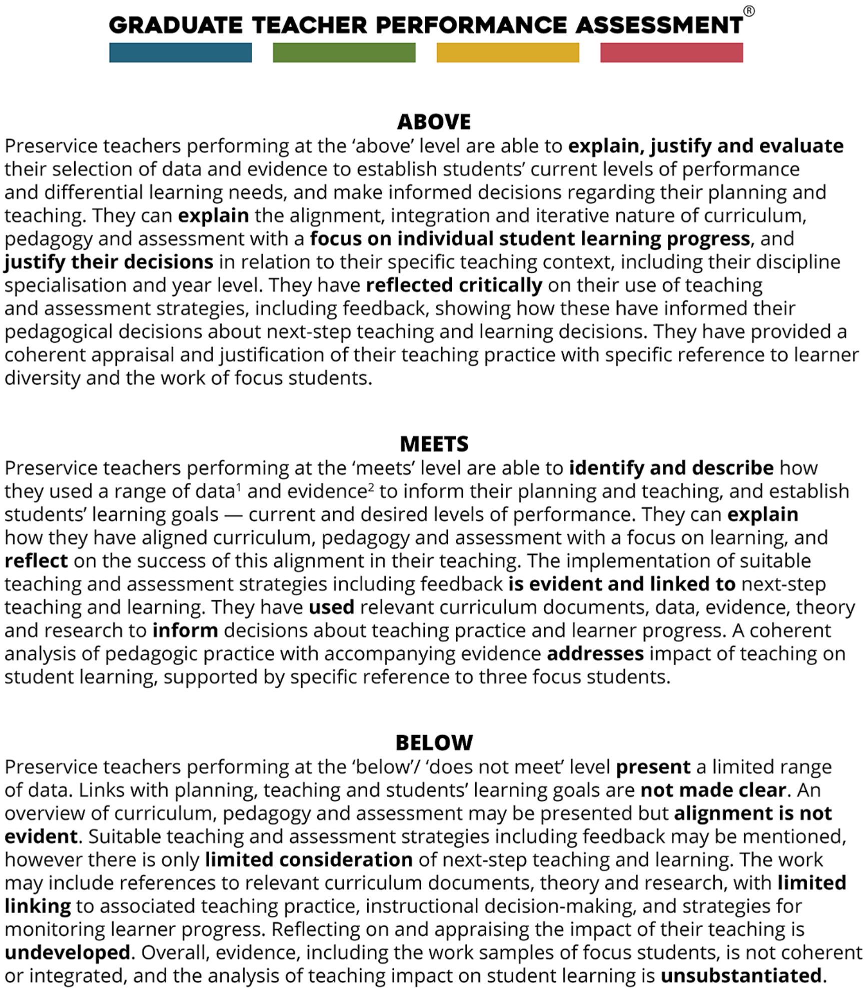
Application of both the DPJM in Study 1 and the pairwise comparison method in Study 2 required the participation of teacher educators who had knowledge of the GTPA content and direct experience with individuals who would undertake the GTPA. There can be no doubt that the trial involved risk-taking on the part of teacher educators (see, for example, Doyle et al., 2021; Lugg et al., 2021). For the first time, they were subjecting their individual judgment processes and ratings to their own scrutiny and that of a larger cross-institution group through a range of paneling processes. The largely private act of arriving at a rating was being de-privatized in the quest to identify a cross-institution benchmark for the GTPA standard.

The application of both the DPJM and the method of pairwise comparisons required careful research design and planning for how selected GTPA samples were to be allocated for scoring to sub-groups (panels) of the participating teacher educators. The creation of stacks of samples that represented a diversity of contexts, the formation of panels of teacher educators to represent the range of universities, and panel discussions on the outcomes of scoring was facilitated by standard-setting experts, including those from ACS Ventures, LLC (Las Vegas, U.S., www.acsventures.com) and the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU).

Study 1: Consistency in judgment and criteria discrimination

In Study 1, the DPJM drew on the collective judgment of 52 teacher educators from the 13 universities that implemented the GTPA in five states and territories of Australia. The participants, who agreed to contribute to the study, were all experienced teacher educators with several years of evaluative experience in assessing academic performance using criteria. At the beginning of the trial however, the criteria for assessing GTPA performances had not been formulated. At this entry point to the trial, there could be no expectation of well-developed, common understandings of either the assessment itself, the scoring rubric, or the expected standard. In addition, the group of teacher educators had no prior knowledge or experience of sharing assessments of scripts outside their own university; they had no history of collaboration in scoring assessments or cross-institutional moderation in teacher education, and they represented a wide range of institutional contexts in metropolitan and rural campuses across the country. Researchers from ILSTE (ACU), and the facilitators designed a process of steps to support teacher educators as they worked through the elements of the DPJM, the initial generation of data from scoring, and the discussions during paneling that were subsequently shared in whole group forums over six days.

The first stage of Study 1, identified as Study 1.1, covered initial review and discussion of the GTPA instrument and a decision aid that included draft PLDs (meets the standard, above, below; see Figure 6.1) as well as the scoring rubric (five defined criteria of *Planning*, *Teaching*, *Assessing*, *Reflecting*, and *Appraising*; see Chapter 4).



GRADUATE TEACHER PERFORMANCE ASSESSMENT®

ABOVE

Preservice teachers performing at the 'above' level are able to **explain, justify and evaluate** their selection of data and evidence to establish students' current levels of performance and differential learning needs, and make informed decisions regarding their planning and teaching. They can **explain** the alignment, integration and iterative nature of curriculum, pedagogy and assessment with a **focus on individual student learning progress**, and **justify their decisions** in relation to their specific teaching context, including their discipline specialisation and year level. They have **reflected critically** on their use of teaching and assessment strategies, including feedback, showing how these have informed their pedagogical decisions about next-step teaching and learning decisions. They have provided a coherent appraisal and justification of their teaching practice with specific reference to learner diversity and the work of focus students.

MEETS

Preservice teachers performing at the 'meets' level are able to **identify and describe** how they used a range of data[1] and evidence[2] to inform their planning and teaching, and establish students' learning goals — current and desired levels of performance. They can **explain** how they have aligned curriculum, pedagogy and assessment with a focus on learning, and **reflect** on the success of this alignment in their teaching. The implementation of suitable teaching and assessment strategies including feedback **is evident and linked to** next-step teaching and learning. They have **used** relevant curriculum documents, data, evidence, theory and research to **inform** decisions about teaching practice and learner progress. A coherent analysis of pedagogic practice with accompanying evidence **addresses** impact of teaching on student learning, supported by specific reference to three focus students.

BELOW

Preservice teachers performing at the 'below'/ 'does not meet' level **present** a limited range of data. Links with planning, teaching and students' learning goals are **not made clear**. An overview of curriculum, pedagogy and assessment may be presented but **alignment is not evident**. Suitable teaching and assessment strategies including feedback may be mentioned, however there is only **limited consideration** of next-step teaching and learning. The work may include references to relevant curriculum documents, theory and research, with **limited linking** to associated teaching practice, instructional decision-making, and strategies for monitoring learner progress. Reflecting on and appraising the impact of their teaching is **undeveloped**. Overall, evidence, including the work samples of focus students, is not coherent or integrated, and the analysis of teaching impact on student learning is **unsubstantiated**.

FIGURE 6.1 Performance level descriptors: Above, Meets, and Below. (Note that this figure represents an artifact used in the 2017 GTPA trial.) Readers interested in the current work on the GTPA and related documentation are advised to contact the two lead authors of this book.

This early stage of judgment with online scoring of GTPA performance samples provided baseline data for an initial assessment of consistency in judgment and criteria discrimination. The second stage of Study 1, identified as Study 1.2, provided additional calibration opportunities and the refinement of judgment processes. This led to improved consistency of judgment and positive changes in criteria discrimination as the teacher educators became more experienced with judging GTPA samples.

Study 1.1: Establishing the judgment process

For the Study 1.1 judgment activity, 32 GTPA performance samples were selected by the research team from the total pool of samples provided by the collective of universities, to represent the full range of teacher education programs and diverse contexts (e.g., across programs, geographic area, subject, teaching year levels) of participating universities. An initial collaborative discussion of the proposed PLDs (Figure 6.1) for assessing a performance as below, meets, or above the GTPA standard was undertaken to begin to probe individual and collective understanding of the expected knowledge and skills to be demonstrated by preservice teachers at the point of entry to the profession. The PLDs shown in Figure 6.1 were revised in consultation with the Collective at key points in the trial; Figure 6.1 does not therefore represent the final version. The discussion was grounded in the examination of samples and the sharing of expertise and expectations of what readiness for the profession means.

The next step involved using the PLDs and the scoring rubric to inform judgment of three illustrative sample performances considered to be of performance quality near to the minimum acceptable level, above the standard, and below the standard. Each of the three samples was examined closely by the group to identify salient features of the performance that aligned to the relevant PLD and the criteria. The purpose of the discussion was to develop a shared understanding of the meaning of the words used in the criteria and how they related to the distinguishing properties of standards at the threshold level of meeting the standard.

Participants were then asked to score each sample against the criteria. This was followed by a structured discussion within the group to articulate an understanding of the level of competence required for licensure to the profession. It was intended to probe the deep structures of the judgment process, facilitated to move from the latent judgment decision to explicit connection to terms that could be used in the rubric and applied in scoring. This understanding was to inform subsequent rating of performances.

Following this discussion, the 52 teacher educators were placed into three panels with participants from the same university systematically allocated to different panels. The 32 GTPA samples were placed into four stacks of eight. One stack of samples was allocated as common to all three panels and the remaining three stacks of sample performances were randomly allocated uniquely to the three panels. The samples were loaded into the online software platform for scoring. Each teacher educator (rater) was asked to read and score 16 performances online to identify the level that best matched each performance with reference to the PLDs and scoring rubric. This was the first application of online cross-institutional moderation in assessing TPAs. This paved the way for the development of the CIM-Online™ [4] platform as discussed in Chapter 7.

Raters completed the scoring task by identifying: the performance with the lowest level of 'meets'; the performance with the highest level of 'does not meet'; and two performances that represented the transition from a level of 'meets' to 'above meets'. The panels of raters then formed small working groups to develop profiles of the level of performance that represented the transition points. Subsequently, the raters were presented with the results of their scoring and a summary of the results from the whole group of participants. This provided the second indication of judgment consistency and an opportunity to discuss the reasons for scoring decisions, that is, to identify the features in the work that were being valued. From this process, an interim consensus recommendation for each transition point was reached and samples illustrative of the standard at the levels 'meets', 'below meets', and 'above meets' were identified.

Study 1.2: Moderation to support judgment consistency

For Study 1.2, 48 teacher educators who had also participated in Study 1.1, came together in a single location over three days to further refine the GTPA judgment process. This time, a total of 108 GTPA performance samples were selected from the total pool of samples, including two training samples. The training included a collaborative review of the PLDs, revised in Study 1.1, and a discussion of the scoring criteria that described the expected knowledge and skills to be demonstrated by preservice teachers. For the purposes of calibration, each teacher educator was tasked with practicing the scoring method by applying the criteria to two GTPA samples. In this task, attention was more sharply focused on the descriptors within each of the five criteria. An important part of this training was to focus on the alignment of the GTPA with the Professional Standards (AITSL, 2011) and specifically, to identify the multiple opportunities provided in the GTPA for demonstrating aspects of the Standards.

Raters were allocated to nine sub-panels with three sub-panels in each of the three panels identified as panels A, B, and C. Sixteen samples were pre-selected to represent a full range of performance quality. From these samples, four groups of four were allocated separately to all raters; panel A only; panel B only; and panel C only. The remaining 90 samples were uniquely allocated across the nine sub-panels so that each teacher educator scored ten samples unique to the sub-panel and eight samples from the common pool. The stacks of 18 samples allocated to each sub-panel were then made available to the raters online for scoring which took place over three days. At the completion of this activity, the score data were analyzed, and the results were reported back to the group. Following discussion of the results, a consensus was reached for an interim policy recommendation that a preservice teacher must satisfy the requirements on at least four criteria to achieve an overall performance score corresponding to meeting the standard. This recommendation is discussed in terms of its consistency with the analyzed judgment data. See the Results section below.

Analytic strategy

The samples were rated using an ordered scale: well below meets, below meets, at meets, above meets, well above meets. This scale was used to provide an overall judgment score for the performance, in addition to a score for each of the five criteria. In reporting the results from statistical analyses, we refer to the criteria specified in the rubric as criterion 1 (*Planning*), criterion 2 (*Teaching*), criterion 3 (*Assessing*), criterion 4 (*Reflecting*), and criterion 5 (*Appraising*). As the primary focus was on consistency in whether a performance was rated as meeting the GTPA standard or not, these scores were converted to a dichotomous scale (1 = meets, 0 = does not meet).

Consistency of judgment was analyzed using a descriptive analysis that computed percentage agreement in rater scores at the criterion level, and an adjusted measure of inter-rater agreement (reliability) obtained from fitting a multi-facet Rasch model (MFRM) to the criterion scores allocated to samples by multiple raters. The MFRM model (Congdon & McQueen, 2000; Linacre, 1994) was used to simultaneously estimate performance quality, criterion difficulty, and rater severity on a single log-linear scale. The maximum likelihood estimates for all parameters in the model were computed using the FACETS Winsteps® software, version 3.80.4 (Linacre, 2018). The goodness-of-fit of the MFRM, in other words how well the model fits the data, was examined using fit statistics in the form of mean-square values and standardized z-statistics.

Results from Study 1

In the first moderation activity in Study 1.1, 52 judges each rated 16 GTPA samples so that a total of 832 judgments were recorded for the 32 performances. Table 6.1

TABLE 6.1 Judgment frequencies at criterion level (does not meet/meets) by total meets score for the sample performances across the five criteria (possible values 0–5). (Note that the percentage of performances that were judged as meeting the standard is not representative of all GTPAs completed in 2017 but reflects the range of quality in performances deliberately selected for Study 1.1.)

*Sum of the five criterion scores on a GTPA sample**	*Overall judgment of a GTPA sample*		*Total judgments*	*Percentage of total*
	Does not meet	*Meets*		*(%)*
0	76	1	77	9
1	57	2	59	7
2	63	9	72	9
3	11	83	94	11
4	0	100	100	12
5	0	430	430	52
Total	207	625	832	100

* A sum of 0 corresponds to all five criteria being judged as unsatisfactory and a sum of 5 corresponds to all five criteria being judged as satisfactory.

shows the number of criteria for which a performance was rated as satisfactory and its association with the overall judgment score allocated by the same rater.

Of the 832 sets of scores, 625 (75%) were judged as meeting the standard overall. Of these 625 scores, 530 (64% of the total) were judged as meeting the standard on either four or five criteria. All samples that were judged as satisfactory on four or five criteria were also judged as meeting the standard overall. For the 94 performances that were judged as satisfactory on three criteria only, 83 (88%) of these were also judged as meeting the standard overall. The judgment that fewer than three criteria were satisfactory for a sample was overwhelmingly associated with a judgment of below the standard overall. These findings raised the question of which three or four criteria were considered most important for a sample judgment decision of meets the standard overall.

Table 6.2 shows that the agreement level for performances scored in Study 1.1 is 82% for the overall judgment and ranged from 76% to 82% for the criterion-level judgments. The lowest levels of agreement are high at 76% for criterion 5 and 78% for criterion 4.

The MFRM model was fitted to the total 4160 rater scores from the 832 judgments of 5 criteria. The criteria related to *Planning*, *Teaching*, and *Assessing* were very familiar to teacher educators and were recognized as the carry forward of what they have characteristically attended to in teacher education. Teacher educators reported widely divergent expectations of what could be included in *Reflecting*, and most reported little experience in assessing *Appraising*. The latter refers to a preservice teacher's practices and use of evidence to discern the impact of their teaching on student learning.

It was not surprising that the estimated difficulty levels were lowest for the criteria relating to *Planning*, *Teaching*, and *Assessing*. The difficulty level was highest for criterion 5 (*Appraising*) which was well-separated from criterion 4 (*Reflecting*) by

TABLE 6.2 Percentage agreement at the criterion level for moderation in Studies 1.1 and 1.2

	Study 1.1	*Study 1.2*
Criterion	*Percentage agreement*	*Percentage agreement*
1	82%	84%
2	81%	85%
3	81%	84%
4	78%	85%
5	76%	79%
Overall	82%	–
Judgments	832	844
GTPA samples	32	106
Min scores/sample	17	4
Max scores/sample	52	48

more than four standard errors; criterion 4 was well-separated from criteria 1–3 by more than five standard errors (separation index is 5.03). The results showed that raters had placed a higher difficulty weight on satisfying criteria 4 (*Reflecting*) and 5 (*Appraising*) relative to satisfying the criteria of *Planning*, *Teaching*, and *Assessing*. The results supported the interim recommendation by teacher educators that an overall score of meets the standard should be awarded to the GTPA if four of the five criteria are satisfied to meet the standard. This would ensure that at least one of the more highly weighted criteria (4 or 5) contributed to the outcome of meeting the standard. Overall rater consistency was moderate with an estimated inter-rater agreement at the criterion level of 0.70.

Study 1.2 included the second calibration and cross-institutional moderation activity. It provided the opportunity for teacher educators to further refine their collective understanding of the constructs being assessed by the GTPA. A second collaborative judgment activity that continued to address construct validity and examine the judgment of performances was critical for improvement in the already high level of rater consistency. In this study, 48 judges each rated a subset of between 16 and 18 GTPA performances at the criterion level so that a total of 844 judgments were recorded for the 106 performances, mentioned earlier. An independent overall judgment was not recorded, as the overall score was computed as meeting the standard if any four of the five criteria were judged as meets. This was the decision reached by the panel of teacher educators following the first moderation activity in Study 1.1.

Descriptive analysis of judgment data from Study 1.2 showed that the percentage agreement in performance had increased for all criteria due to the intensive calibration activities over three days, with the largest improvement occurring for the criterion of *Reflecting* (see Table 6.2). This demonstrated the improvement in construct validity for the attribute of reflecting. Results from the MFRM model showed that the difficulty level for the criterion of *Reflecting* came into line with the three criteria of *Planning*, *Teaching*, and *Assessing*. The inter-rater agreement estimated from the MFRM model increased from 0.70 in Study 1.1 to 0.78 in this second moderation activity. This confirmed the potential of calibration in a collaborative environment to improve the inter-rater agreement. Readers are invited to see Chapter 7 for a discussion on the association between participation in calibration, cross-institutional moderation, and reliability.

Measures of rater agreement were consistently lower for the criterion of self-appraisal of teaching relative to the other four criteria indicating that construct validity for the attribute of *Appraising* could be improved at this early stage of GTPA implementation. Rater consistency was lower for GTPA performances that were close to the threshold of meeting the standard, demonstrating that ongoing training will be required to improve consistency of judgment for performances at this level. While these insights were important for improving future implementation of the GTPA, the approach outlined above provides evidence of the validity of the instrument such that the underlying constructs are recognizable to experts in the field and able to be judged with consistency.

Study 1 demonstrated the successful application of the DPJM for meeting objective (1) identifying the characteristics of performance to be recognizable as competent (ready) for classroom practice, and for deriving a description of the minimum acceptable performance level for the agreed standard. The study also demonstrated the reliability of the instrument and the value of calibrating judgment for improving consistency of judgment. The improvement in judgment dependability was achieved through a combination of sustained talk, interaction among teacher educators, and customized artifacts, including textual resources.

Absent from Study 1 was the identification of GTPA performance samples that exemplified the minimum acceptable level of performance at which a sample meets the standard. The availability of exemplars that demonstrate threshold performances against the standard and across different contexts can be a critical aid for establishing and maintaining consistency in judgment for a complex performance assessment with multiple interlocking criteria such as the GTPA (Sadler, 2009). Building on the work achieved in Study 1, the purpose of Study 2 was to design and implement a rigorous process to identify exemplars to support teacher educators in decision-making for judgment of the GTPA – an outcome that would contribute to the sustainability of consistency in judgment against the standard through future iterations of cross-institutional moderation.

Study 2: Setting the standard with pairwise comparison methodology

As discussed above, Study 2 combined two adjacent judgment methodologies in the approach to standard setting. In the first part of this study, identified as Study 2.1, a pairwise comparison method was applied in which judgment data were generated by participating teacher educators and analyzed to derive the order of GTPA performance samples from highest to lowest. From this ordered list of samples, a subset was selected to represent performances from across the range of the scale. The second part of the study, identified as Study 2.2, drew on the combined expertise of the most experienced participating teacher educators to identify the adjacent samples from the subset, between which the transition from meeting and not meeting the standard was observed. This was the rigorous process from which exemplars of the minimum acceptable performance level for the standard could be identified.

Study 2.1: Pairwise comparison of GTPA samples

From the total pool of GTPA samples submitted for the trial, a subset of 50 GTPA samples was selected for use in Study 2. The samples were chosen to represent a well-distributed range of performance, across a variety of contexts, but also with a concentration of samples near the expected threshold for meeting the standard. From the teacher educators who contributed to Study 1, 43 were invited to participate in Study 2. They were each presented with ten pairs of GTPA performance samples online and were asked to make a comparison of the samples in a pair, using the rubric for the five

GTPA criteria, to select the highest performing sample. These teacher educators were instructed to "compare each pair and decide, on balance of the evidence, which of the performances demonstrates more advanced understandings, skills and knowledge with respect to the criteria provided" (Wyatt-Smith et al., 2020, p. 72). Of the 43 invited participants, 36 completed the ten allocated paired comparisons.

The generation of 430 pairs from 50 different samples (comprising ten pairs for each of 43 judges) and the design for how pairs of samples would be allocated to judges, required careful consideration. To reduce the time taken for a judge to read all 20 samples, the same sample was included in four different pairs of the ten that were allocated to a judge. This could be achieved using a pair-generation algorithm that incorporated parameters reflecting this decision as well as the specification of the minimum number of times that each sample was to be included in a pair.

The ten sample pairs allocated to a judge were presented online so that each pair of samples appeared on the screen side-by-side. The judge was able to select the sample that they understood to be the superior performance on each of five criteria. The submitted responses were stored in a central database that recorded the binary decision score against each criterion. Results from analysis of these data placed each of the 50 samples in order of performance on a derived scale. The order of the performance samples could be used to select samples for inclusion in the subsequent pairwise comparison activity for identifying the transition to meeting the standard at the threshold. This was the focus of Study 2.2.

Study 2.2: Setting the standard with exemplars of the minimum acceptable level

The expert panel who participated in Study 2.2 included 16 experienced teacher educators selected from the pool of judges who had also participated in Study 2.1, with representation from all universities in the GTPA Collective. A subset of ten samples was selected from those that were judged with consistency in Study 2.1 and were well separated by relative performance. Samples were excluded from selection if they were overly difficult to judge in Study 2.1 and had comparatively pronounced differences in criteria judgments. The focus of the study was on identifying the sample that exemplified the minimum acceptable level of performance at the threshold of the standard.

The expert teacher educators were presented with the same ten samples, ordered relatively from lowest to highest standard, using the online Pairwise Comparison Application. The samples were ordered according to their scale location derived in Study 2.1. Instructions to the panel were:

> Please read and become familiar with the performance to the right. Make an on-balance judgement about whether the performance meets or exceeds the minimum standard you would expect from a graduate teacher. If you think that performance meets or exceeds the minimum standard, select that performance.
>
> *(Wyatt-Smith et al., 2020, p. 74)*

An alternative option was available for selection if the performance was considered to be below the standard.

Analytic strategy

The binary data generated from pairwise sample comparisons in Study 2.1 were analyzed using the Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959). This model estimated relative scale locations for each of the 50 GTPA samples. To summarize the internal consistency of the pairwise comparisons, a performance separation index was produced overall and separately for each criterion. For applications in which samples cannot all be compared against each other, the scaling algorithm in the model adjusts the sample locations relative to other performances that a sample has been compared against. This was important for Study 2.1 where it was not time-efficient to compare each sample with all 49 other samples. The number of times for which each of the samples was judged against a different sample varied from 9 to 25.

As mentioned above, the ordered scale locations estimated for the 50 samples included in Study 2.1 were used to select the subset of ten samples for Study 2.2. Analysis of the binary decisions by the panel of 16 expert judges generated statistical estimates for the locations of the ten samples on a logit scale, including the locations of the samples relative to each of the five criteria. Additionally, information from the analysis could be used to derive a threshold that represented the position on the scale above which sample performances had met the standard.

The threshold level was displayed visually as a line in a graph that also showed the locations of exemplars relative to each of the five criteria. This enabled the panel to compare the profiles of strengths and weaknesses on the criteria against the threshold. The panel was presented with this visual representation of the results from the analysis (see Figure 6.2) to provide the opportunity for discussion to determine whether there was consensus agreement on the threshold for the standard and the sample performance that demonstrated the minimum acceptable level for entry to the profession.

Study 2: Results

The BTL model was used to estimate scale locations for each of the 50 samples included in Study 2.1, and to estimate separation indices for the five criteria. Internal consistency in the pairwise comparisons was high (0.95). The criterion level separation indices were 0.77 (*Planning*), 0.75 (*Teaching*), 0.80 (*Assessing*), 0.81 (*Reflecting*), and 0.83 (*Appraising*). The information on relative scale locations and separation indices generated from this analysis were used to select ten samples that were well separated on the scale for inclusion in Study 2.2.

Following pairwise comparisons of the ten samples in Study 2.2, Table 6.3 (table 6 in Wyatt-Smith et al., 2020, p.78) shows the number and percentage of occasions on which panel members indicated that the performance of a GTPA sample was

TABLE 6.3 Numbers (N) and percentages of panelists who scored sample performances above the threshold of meeting the standard

Sample ID	*No. of comparisons*	*N*	*Location*	*Percentage (%)*
201700171	16	1	−3.472	6
201700121	16	0	−1.294	0
201700114	16	7	−0.799	44
201700192	16	10	−0.447	63
201700209	16	16	−0.173	100
201700076	16	15	0.273	94
201700180	16	15	1.075	94
201700015	16	16	1.242	100
201700316	16	16	1.486	100
201700005	16	16	2.213	100

above the threshold of the standard. The samples (represented by unique numeric identifiers) are shown in order of scale locations (estimated using the BTL model), indicating the order of performance from lowest to highest down the column. The panel of judges agreed that the threshold location was between the third and fourth samples from the top of Table 6.3. This corresponded to a 19% shift in judge agreement, from 44% for sample "201700114" to 63% for sample "201700192", that the sample performance was above the threshold of meeting the standard. For sample "201700209" at location −0.173 and samples located higher on the scale, there was almost unanimous agreement that these samples were performing above the threshold.

Figure 6.2 (Figure 2 in Wyatt-Smith et al., 2020, p.76) visualizes the location of each sample on the logit scale. This graph shows the scale locations of sample performances (represented on the vertical axis) against the criteria (represented on

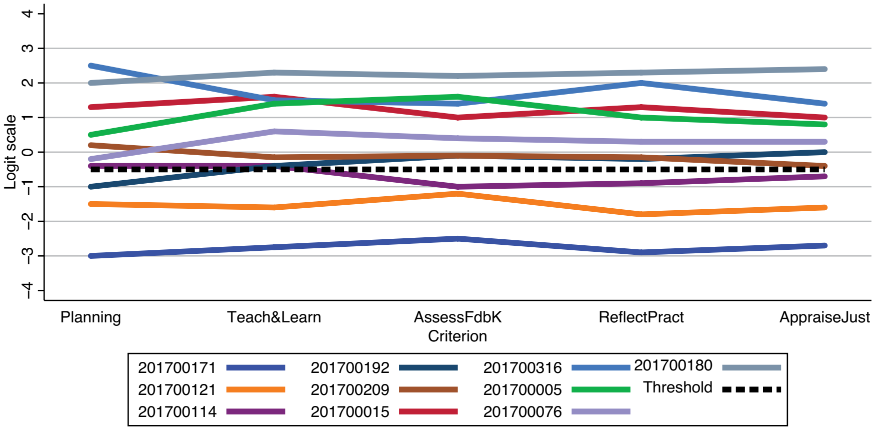


FIGURE 6.2 Profile representation of performances relative to threshold

the horizontal axis), joining the locations of the five criteria for each sample. The solid thick line corresponds to the threshold level as defined above, located between samples "201700114" and "201700192". Figure 6.2 shows that these two samples are most separated by judgments on the two criteria of *Assessing* and *Reflecting*.

In determining the threshold level for the standard, there was some level of disagreement within the panel, as would be expected. However, discussions in the panel meeting that followed the standard-setting exercise, led to a consensus that the threshold level should be set between the third and fourth exemplars in order. There was agreement that it would not be defensible to set the threshold position lower. The panel subsequently recommended that the two exemplars selected in Study 2.2 should be used as threshold exemplars in practice going forward. The implication was that performances that sit above "2017000192" would be deemed to meet the minimum acceptable level of the standard.

As mentioned earlier in this chapter, exemplars of performance can be used to illustrate how the characteristics of quality have been applied in judgment. When accompanied by a description of features relating to the various criteria, the exemplars can guide the on-balance judgment of the level of each performance. For performances that are very near the threshold of the standard, judgment is inherently more difficult to make. Explicit guidance on decision-making mechanisms in cases in which a performance is considered close to the threshold level is critical for achieving consistency in judgments.

The cumulative value of two methodologies

The purpose of the two studies undertaken in the 2017 GTPA trial was to establish judgment consistency, reliability of the instrument, and the threshold for setting the standard of graduate readiness to enter the teaching workforce. Three questions relevant to setting the standard were:

> *What is involved in establishing the standard? Can we reach an agreement on the standard? How do we make the standard visible to those in the profession?*

The sequential objectives related to addressing these questions were to (1) identify the characteristics of performance to be recognizable as competent (ready) for classroom practice, (2) determine the minimum acceptable performance level for the agreed standard, and (3) identify illustrative examples of performances at the threshold of meeting the established standard. These objectives could only be achieved through the collective effort of a representation of teacher educators who formed a guild of GTPA assessors. The participation of a large group of teacher educators enabled the design of progressive studies to establish judgment consistency and reliability of the instrument within the overarching studies for achieving the standard-setting objectives.

Achieving objectives 1–2 for a complex integrated task such as the GTPA, incorporating the expertise of as many as 80 professionals, is not straightforward.

Determining the agreed characteristics of performance by preservice teachers that are seen to represent competence in teaching (objective 1) and then deriving a provisional shared understanding of a recommended level of performance for meeting the standard (preliminary to objective 2), could be best managed and achieved through a multi-stage judgmental process. A methodology that combined the DPJM and the analytic judgment method was identified to be most suited to meeting these objectives for a complex performance assessment, as it provided the framework to incorporate both discussion and empirical evidence into judgments while permitting experts to make their recommendations before reaching the final agreed collective determinations. Moderation of performance samples was a key component of the process which provided the means to monitor and establish judgment consistency at the same time.

While this methodology used in Study 1 was successful in achieving objective 1 and providing the groundwork for objective 2 in describing the characteristics of the minimum acceptable performance level for the standard, it could not determine what the minimum acceptable performance at the threshold for the standard would look like. To provide exemplars that illustrated GTPA performances at the agreed threshold of the standard required an additional methodology that had been designed specifically for this purpose. One such approach that had been used successfully to identify exemplars of student performance for the common assessment grade scale in Australian schools is the method of pairwise comparison. This method was readily adaptable for identifying GTPA exemplars of the standard and was appropriate for completing objective 2 and achieving objective 3 in a second study (Study 2). Both studies were grounded in the evaluative expertise of teacher educators in the human judgment of GTPA performances and the decision-making processes that followed the feedback of results from analysis of the combined judgment data.

The two methodologies together provided the mechanisms to achieve all three objectives for setting the GTPA standard. The multi-stage judgmental methodology was critical for reaching an agreement by the profession on what readiness to teach in the classroom was recognized to be. It was also important for establishing the baseline for judgment consistency and for mobilizing the mechanisms for improving and sustaining judgment consistency across institutions. However, this methodology on its own did not include the study design or infrastructure necessary for objectively identifying illustrative exemplars of performance at the threshold of the standard. These features were provided by the complementary method of pairwise comparison. Both methodologies applied in Study 1 and Study 2 were necessary to *make the standard visible to those in the profession.*

Summary

As the GTPA was first implemented in participating universities in 2017, it was necessary to establish the validity of the instrument and reliability of judgment across multiple institutions against an agreed standard. A multi-stage judgmental process, with participation from cross-institution teacher educators, was adopted to ensure that the constructs for assessment and the standard to be achieved were informed and well understood by the teacher educators who were involved in delivering the ITE programs and judging performance on the GTPA. The teacher educators needed to bring their evaluative expertise to the table and be willing to fully participate in the activities designed to derive the minimum acceptable level of GTPA performance. This was achieved through lengthy discussion and human interaction over a one-year period including eight days of workshop sessions.

The judgmental process took place in four parts. Study 1.1 and Study 1.2 incorporated collaborative activities to ensure construct validity for the GTPA instrument and to support decisions in assessing performance leading to consistency in judgment. Study 2.1 and Study 2.2 incorporated pairwise comparison activities that identified ordered performances, including exemplars of the minimum acceptable performance level for the standard. The methodologies adopted in both studies were layered with opportunities for valued human judgment to intertwine with quantitative data analysis at key points in the process. Results from data analysis were combined with further human interaction to guide decision-making.

The standard-setting activities provided the first opportunity for teacher educators from diverse and dispersed universities to externally moderate GTPA performance decisions. The online system for scoring, data collection, analysis, and immediate reporting back to the panel for discussion and decision-making used in 2017 was the beginning of the design of customized infrastructure for sustaining GTPA judgment and reporting activities. Chapters 3 and 7 propose that the development of a common assessment instrument is simply the entry point for enabling teacher educators' engagement with TPAs and does not on its own guarantee ongoing fidelity of implementation and reliability in judgment. The approaches taken in this chapter provided the underpinnings for developing the infrastructure for ensuring the sustainability of ongoing cross-institutional moderation, judgment consistency, and application of the agreed standard beyond 2017. The enhanced digital infrastructure now in use by the GTPA Collective of participating universities is described in Chapter 7.

The digital infrastructure developed by ILSTE (ACU) researchers is known as *Evidence for Quality in Initial Teacher Education* (EQuITE) and includes the cross-institutional moderation online system, GTPA CIM-Online™ (see Chapter 7 for details). The system incorporates GTPA sample submission by participating universities and processes for checking the deidentification of the samples. The system architecture has in-built features that guarantee data security and privacy of data submitted by different universities. For the purposes of cross-institutional moderation, system features include a portal that enables calibration for participating teacher educators (judges) before commencing online moderation, an algorithm for designing the systematic (partially random) allocation of samples to judges from different universities, and a secure portal for submitting online judgments of allocated samples (Chapter 7). Chapter 8 presents the data analytic methods and visualization of results used to provide feedback to universities to inform decisions related to ITE curriculum review and program renewal. This feedback is presented in the form of confidential reports that are generated separately for each university by the digital EQuITE system. Readers are encouraged to turn the page for more insights into the system that sustains these ongoing practices.

Notes

[1] Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

[2] In Australia, teacher registration and the accreditation of teacher education programs are the responsibility of state regulatory authorities.

[3] The study required ethics approval including permissions from participating preservice teachers where their samples would be used for research purposes.

[4] Acknowledgment: The online model of cross-institutional moderation (CIM-Online™) was conceptualized and developed in the Institute for Learning Sciences and Teacher Education, Australian Catholic University. For a discussion of CIM-Online™, readers are advised to also see Wyatt-Smith and Adie (2021).

References

Australian Institute for Teaching and School Leadership [AITSL]. (2011). *Australian professional standards for teachers*. https://www.aitsl.edu.au/teach/understand-the-teacher-standards/how-the-standards-are-organised

Australian Institute for Teaching and School Leadership [AITSL]. (2018). *Accreditation of initial teacher education programs in Australia: Standards and procedures*. https://www.aitsl.edu.au/tools-resources/resource/accreditation-of-initial-teacher-education-programs-in-australia---standards-and-procedures

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs, I. The method of paired comparisons. *Biometrika*, *39*(3–4), 324–345. https://doi.org/10.2307/2334029

Bramley, T., & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education*, *25*(3), 293–317. https://doi.org/10.1080/02671522.2010.498147

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, *73*(2), 163–178.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgement-in-context: Analysing how teachers evaluate students' writing. *Educational Research & Evaluation*, *13*(5), 401–434. https://doi.org/10.1080/13803610701728311

Craven, G., Beswick, K., Fleming, J., Fletcher, T., Green, M., Jensen, B., Leinonen, E. & Rickards, F. (2014). *Action now: Classroom ready teachers*. https://docs.education.gov.au/documents/action-now-classroom-ready-teachers-report-0

Doyle, T., Evans, N., & Salter, P. (2021). Opportunities and tensions in the experiences of collaborative professionalism during the enactment of the GTPA. In C. Wyatt-Smith, L. Adie, & J. Nuttall (Eds.), *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration* (pp. 81–94). Springer.

Freebody, P., & Wyatt-Smith, C. (2004). The assessment of literacy: Working the zone between "system" and "site" validity. *Journal of Educational Enquiry* *5*(2), 30–49.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th Ed., pp. 433–470). Praeger.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245–270. https://doi.org/10.1080/02671520500193744

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*(2), 1–20.

Heldsinger, S., & Humphry, S. (2013). Using calibrated exemplars in the teacher-assessment of writing: An empirical study. *Educational Research*, *55*(3), 219–235. https://doi.org/10.1080/00131881.2013.825159

Humphry, S., Heldsinger, S., & Dawkins, S. (2017). A two-stage assessment method for assessing oral language in early childhood. *Australian Journal of Education*, *61*(2), 1–17. https://doi.org/10.1177/0004944117712777

Humphry, S. M., & McGrane, J. A. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *The Australian Educational Researcher*, *42*(4), 443–460. https://doi.org/10.1007/s13384-014-0168-6

Linacre, J. M. (1994). *Many-Facet Rasch measurement.* (2nd Ed.). MESA Press.

Linacre, J. M. (2018). A user's guide to FACETS Rasch-model computer programs. Program Manual 3.81.0. http://www.winsteps.com/facetman/titlepage.htm

Luce, R. D. (1959). *Individual choice behaviours: A theoretical analysis.* J. Wiley.

Lugg, A., Lang, C., Weller, J., & Carr, N. (2021). Collaboration in a context of accountability: Cultural change in teacher educator practice across university boundaries. In C. Wyatt-Smith, L. Adie, & J. Nuttall (Eds.), *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration* (pp. 95–114). Springer.

Plake, B. S., & Hambleton, R. K. (2001). The analytic judgement method for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 283–312). Erlbaum.

Plake, B. S., Hambleton, R. K., & Jaegar, R. J. (1997). A new standard-setting method for performance assessments: The dominant profile judgement method and some field-test results. *Educational and Psychological Measurement*, *57*(3), 400–411. https://doi.org/10.1177/0013164497057003002

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144.

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, *34*(7), 807–826.

Smith, C. (1989). *A study of standards specifications in English* [Unpublished Master's thesis]. University of Queensland.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*(4), 273–286. http://dx.doi.org.ezproxy2.acu.edu.au/10.1037/h0070288

Wyatt-Smith, C., & Adie, L. (2021). Introducing a new model for online cross-institutional moderation. In C. Wyatt-Smith, L. Adie, & J. Nuttall (Eds.), *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration* (pp. 39–58). Springer.

Wyatt-Smith, C., Humphry, S., Adie, L., & Colbert, P. (2020). The application of pairwise comparisons to form scaled exemplars as a basis for setting and exemplifying standards in teacher education. *Assessment in Education: Principles, Policy & Practice*, *27*(1), 65–86. https://doi.org/10.1080/0969594X.2020.1712326

7

THE DESIGN FEATURES OF CROSS-INSTITUTIONAL MODERATION FOR DEMONSTRATING COMPARABILITY

Building a nationally sustainable model

Introduction

In this chapter, we outline the key elements in a new framework for cross-institutional moderation (CIM) that combines statistic and social elements of moderation. It has been conceptualized and designed for the Graduate Teacher Performance Assessment (GTPA®).[1] We propose that this framework is potentially applicable not only to other teaching performance assessments (TPAs) internationally, but also to authentic, complex, performance assessments in other professions. CIM, as referred to in this chapter, involves assessors scoring authentic samples provided by multiple universities. This is a blind review process with samples being fully de-identified, including the removal of the original score provided by the host institution. The term CIM-Online™ [2] refers to the use of digital technologies and online scoring systems to record and collate judgment decisions using an established standard (see Chapter 6).

CIM optimizes the potential of TPAs. Our position is based on two propositions. First, there must be a focus on CIM and cross-TPA moderation to avoid the risk of producing a two-tiered or a multi-tiered system of teacher education in the country. CIM is central to university efforts to monitor movement in a standard over time, both within and across programs. A nation can move to introduce TPAs to leverage improvement in teacher education. Expert panels can be set up to review TPAs and establish conditions under which they can be endorsed. However, without a concurrent focus on CIM, a TPA can become 'just another assessment' without any connection to a prepared-to-teach standard. By focusing on the TPA as an instrument, rather than TPA CIM, the opportunity to have a serious conversation about the expected quality of teaching graduates entering the profession could be missed. In addition to endorsing the instrument, we argue that a common standard be

DOI: 10.4324/9780429318504-9

applied across TPAs. This requires the development of quality assurance systems and processes, including CIM.

Our second proposition is that a national decision to regulate the number of TPAs (say to one) is not the answer to improving the quality of teacher education. Rather, the uptake of a data-informed approach to advancing the quality of teacher education requires new thinking about benchmarking and the principles and practices of cross-TPA moderation. Three vexed questions play out in considering the importance of CIM in lifting the quality and effectiveness of initial teacher education (ITE): (1) *Is it sufficient to undertake intra-university moderation only, that is within a single university in one TPA?* (2) *What is the number of institutions necessary for undertaking CIM in a single TPA?* (3) *What methodologies are suitable for benchmarking across TPAs?* The complexity of addressing what happens when the evidence base is different across TPAs is addressed in the final chapter.

The GTPA (see Chapter 4) was designed and implemented with the intent to (1) be a summative assessment of graduate competence, (2) provide data on program performance that teacher educators could use to inform curriculum review and program renewal, and (3) produce large-scale data on the quality of teacher education in Australia. CIM was identified as integral to each of the three purposes. There were five steps that we took in our approach to CIM:

Step 1: Learning from research, practice, and literature.
Step 2: Making decisions about fitness-for-purpose and the approach to moderation, now extended to the use of digital technologies in GTPA CIM-Online™.
Step 3: Designing to build dependability, including the development of principles of fidelity of implementation of the GTPA, the material artifacts needed to support scoring, and scorer training and calibration to apply the standard to undertake moderation.
Step 4: Training to read and interpret the reports from the analyses of the data produced from moderation.
Step 5: Introducing a feedback loop into initial teacher education to use the data for program improvement.

In this chapter, we provide details on Steps 1–3 and address Steps 4–5 in the following chapter on the feedback loop that is established through CIM. We present a case for the centrality of standards-referenced online CIM practices in building the dependability and defensibility of judgments, and the confidence of teacher educators, preservice teachers, and the public in the quality of graduates entering the teaching profession.

Step 1: Learning from research, practice, and literature

Internationally, where moderation occurs in examinations and schooling systems and in higher education, generally speaking, it is regarded as an “approach to agreeing, assuring and checking standards” (Bloxham et al., 2016, p. 638). While there is no single, agreed approach to moderation in these contexts, typically it is associated with quality assurance and sustaining public confidence in the integrity of the assessment system (Crimmins et al., 2016; Klenowski & Wyatt-Smith, 2010). Approaches to moderation include social moderation where scorers come together to review student work samples, expert moderation where the evaluative experience of a nominated expert is used to make the final judgment decision, and statistical moderation that involves scaling of scored performances. In practice, there can be combinations of these forms of moderation. Evident in both social and statistical approaches in several countries is the move to online moderation processes (e.g., United Kingdom: Pobble (n.d.); South Africa: Rajamany et al., 2020; Australia: Adie, 2013; Tarricone & Newhouse, 2016; Wyatt-Smith & Adie, 2021; and across countries: Gilmore et al., 2020).

It is fair to say that statistical moderation has been the preferred approach in large-scale testing and measurement where reliability has been the priority. Statistical moderation, used in high-stakes examination contexts, reflects a belief that human judgment can be subject to bias, random errors, and even whimsey. In statistical moderation, assessment grades or scores are calibrated or adjusted based on statistical comparison with other assessments or results (Crisp, 2017; Williamson, 2016). Several factors such as the severity and lenience of scorers can be identified in these processes.

Social moderation, by far the less prominent in education systems internationally, has been embraced within some jurisdictions to foster discussions of quality and characteristics of the work assessed against a stated standard. Junior cycle in the Republic of Ireland, and education systems in New Zealand, Australia, and some provinces in Canada, are examples of where social moderation is practiced. This involves teachers meeting to share work samples and discuss the application of standards in their judgments. In some contexts, moderation in universities involves staff meetings to review grades and samples and adjust grades as required before reporting to Faculty Boards. Drawing on our experiences of social moderation systems in schooling and higher education (Adie, 2013; Wyatt-Smith & Adie, 2021, in press; Wyatt-Smith & Colbert, 2014), we adopt the position that moderation is cultural and historical practice and can vary across sites. Broadly speaking, moderation requires that participants take up particular identities and practices.

Demonstrating the dependability of judgments is a well-recognized feature in assessment literature. Dependability of judgments has been located at "the intersection of reliability and content validity" (Wiliam, 1994, p. 18). Harlen (2004) suggested five actions that could be used to investigate dependability: "the specification of the tasks; the specification of the criteria; training; moderation; and the development of an 'assessment community' within the school allied to increased confidence in the professional judgment of teachers" (p. 28). Concerns about dependability have been associated with the influence of bias and varying interpretations of the meaning of criteria and standards, especially when these are written as verbal descriptors (Harlen, 2004).

Literature on moderation, especially social moderation, similarly identifies a range of concerns regarding scorer bias, reliability, rater and inter-rater consistency in applying the standard across large populations, and costs including for travel and staff time. Added to this is the influence of professional identities on the talk and interactions that occur during moderation and the impact of these on judgment decisions (Adie, 2013; Estyn, 2016; Hipkins & Robertson, 2012). The key question for social moderation practices in universities and schools is: What evidence can be produced from social agreement to demonstrate reliability and comparability of judgments within a program, across programs, and across campuses? Building on Wiliam's (1994) and Harlen's (2004) discussions on the dependability of teacher summative assessments, we also explore a new question in teacher education: *Under what conditions can teacher educator judgments of TPAs be made dependable?*

As elaborated in the remainder of this chapter, our approach in the GTPA combines elements of social and statistical moderation in a hybrid online cross-institutional model (CIM-Online™). In going to scale in a consortium of universities (hereafter referred to as the Collective), our aim is to take up the affordances of digital technologies within a new sustainable model for moderation. Our approach is distinctive in how it combines social, statistical, and online processes to enable synchronous and asynchronous interactions. As used in the GTPA, the term online cross-institutional moderation (CIM-Online™) refers to social interactions and digital processes designed to (1) support calibration training for raters, (2) submit candidate work samples, (3) review and score samples via an online platform, (4) store results in a customized data warehouse for analysis, and (5) produce and distribute confidential reports to each participating university. In online moderation, teacher educators' roles as raters include blind review and scoring of authentic samples presented in virtual stacks (described below).

In these ways, we take up the invitation by Bloxham et al. (2016) to advance "an important conversation regarding the 'point' of moderation, which is most clearly understood when practices move beyond accountability to inform teaching and enhance student learning opportunities" (p. 650). Following completion of scoring, data are analyzed and reports are prepared for each participating university. They include information about how the standard has been applied in each ITE program as well as an endorsement of scoring decisions within each institution. Information is also provided about program characteristics and performance using data provided by participating universities. This linking of moderation, scoring, and reporting includes dialogue between data analysts and teacher educators in interpreting the meaning of the reports as it relates to their programs. The dialogue in moderation events opens the data for teacher educator action in curriculum review and program renewal. This stance reflects the authors' view that teacher educators are best placed to engage with, and ascribe meaning to, the data and in so doing, enact evidence-informed pedagogy. The reports are not the end goal.

The reconceptualization of moderation in the GTPA project builds on the earlier discussion in Chapter 3 regarding teacher educators' work at the intersection of system and site requirements for TPAs. Teacher educators' judgments play a critical role at this intersection in terms of how data are used for summative (reporting) purposes and formative (improvement) purposes (see Chapter 8). At the system level, the priorities are to assure that a common standard is being applied across teacher education providers. The site-specific priorities include using the reported results from moderation for reviewing the effectiveness of teacher education programs and curriculum, taking account of community needs and expectations (see Chapter 9 for teacher education voices in this process).

Step 2: Making decisions about fitness-for-purpose and the approach to moderation

Decisions about the form moderation would take varied across three main phases of TPA development in the project (see Figure 7.1 and Table 7.1). The phases were identified as

Phase 1: Validation of the TPA and accompanying scoring rubric and initial standard setting (discussed in Chapter 6). An outcome of Phase 1 standard-setting moderation activities was a scale of performances and the delineation of a level for Meets and Does Not Meet. This was a standard which had not previously existed. This phase established the fidelity principles for Phase 2 implementation of the GTPA across various teacher education sites nationally, and for the design of customized digital architecture to ensure sustainability of the practice in Phase 3.

Phase 2: Intra-university moderation. Each university grades and moderates their own samples taking account of their university assessment policy and using the GTPA-agreed standard and criteria. Teacher educators also

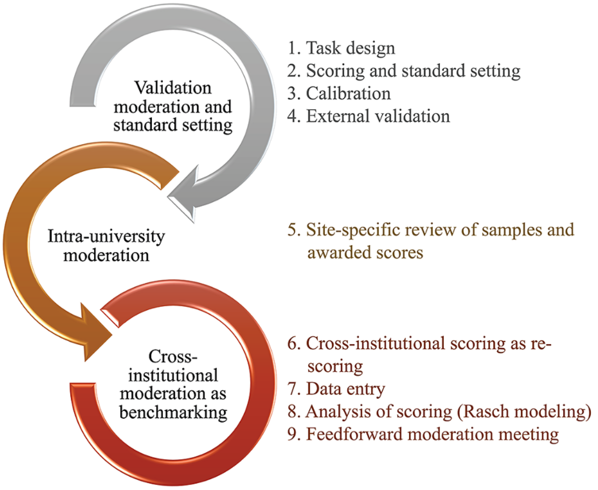

Validation moderation and standard setting
1. Task design
2. Scoring and standard setting
3. Calibration
4. External validation
Intra-university moderation
5. Site-specific review of samples and awarded scores
Cross-institutional moderation as benchmarking
6. Cross-institutional scoring as re-scoring
7. Data entry
8. Analysis of scoring (Rasch modeling)
9. Feedforward moderation meeting

FIGURE 7.1 Moderation across the phases of TPA development and implementation

TABLE 7.1 Moderation as embedded in GTPA activities and processes

Developmental phase of GTPA	*GTPA activities*	*Moderation focus*	*Example moderation activities*
1. Validation and standard setting (For discussion of these activities, see Chapters 4 and 6)	i. Task design	Validation	- Expert review of GTPA for authenticity against professional standards
	ii. Scoring and standard setting		- Expert group discussion of scoring rubric - Applying the rubric in judging samples - Analyzing rater consistency and discrimination among criteria (analytic judgment) - Online pairwise comparison, placing of samples on a scale; internal consistency of raters - Performance profiles relative to threshold (see Chapter 6, Figure 6.2) - Exemplification of the standard
	iii. Calibration		- Workshop activities to develop a shared understanding of the GTPA and rubric
	iv. External validation		- Expert panel review and endorsement
2. Validated TPA	v. Site-specific implementation and assessment	Intra-university moderation	- Consistency of judgments at the program level - Selection of samples for cross-institutional moderation (CIM-Online™)

(*Continued*)

TABLE 7.1 (Continued)

Developmental phase of GTPA	*GTPA activities*	*Moderation focus*	*Example moderation activities*
3. CIM-Online™ for benchmarking	vi. Cross-institutional scoring (re-scoring) and data entry	Reliability of judgments using the standard	- In-person and online calibration activities - Blind review of de-identified samples - Quality range of samples - Overall judgment (Meets/Does Not Meet) - Judgment at criterion level
	vii. Analysis of scoring		- Rasch modeling - Demonstrated consistency in the application of the standard - Rater severity and lenience
	viii. Feedforward moderation meeting	Effectiveness of intra-moderation practices	- Customized reporting: Feedback of judgment decisions and feedforward into curriculum review and program renewal

select and submit samples for GTPA CIM-Online™, administered by the Institute for Learning Sciences and Teacher Education (ILSTE). In addition to submission of samples, the university provides confirmation that the assessment has been implemented in accordance with the GTPA established conditions of fidelity (Adie & Wyatt-Smith, 2020) discussed below, and that internal moderation has been completed. This provides confidence that the submitted samples show the standard as applied within the university and for each program involved in CIM-Online™.

Phase 3: CIM-Online™ as a form of benchmarking. This stage required the development of new systems and digital infrastructure to monitor the application of the standard over time.

A priority of the moderation design for Phases 1 and 3 was to connect teacher educators across the country. Online technologies were therefore essential. Linking teacher educators matters in a context where teacher education quality is a national priority.

Phase 2 processes rested with each university. Our decision-making was focused on Phases 1 and 3 where we sought information about the components necessary to implement moderation in-person and online. This included information about:

1. Digital infrastructure and relevant platforms/software to support moderation.
2. Method and means for recording, storing, and submitting confidential outcomes, including those from individuals and groups of scorers.
3. Data upload and encryption processes and protocols for ensuring privacy and secure transmission of (1) samples to be scored, (2) cohort results and demographic information used in analysis, and (3) customized reports of analyzed data.

We considered the judgment and moderation methodologies best suited to enact Phases 1 and 3. Decisions were related to:

1. Judgment methodologies suited to the purpose of standard setting or that of ongoing moderation activities (see Chapter 6 for a discussion of judgment methodologies used in standard setting).
2. Moderation methodologies and protocols that accommodate the location and scale of moderation activities, as well as generate data necessary to measure the reliability of scoring.
3. A research-informed approach to discerning the optimum ratio of samples to scorers for achieving credibility in the moderation process (i.e., the minimum number of judgments per sample). This reflects how the accepted level of uncertainty in estimating reliability in implementing moderation in teacher education is a vital consideration. The nature and function of virtual stacks of samples were part of these decisions. For example:

   a. What are the benefits of the same sample appearing in different stacks for judgment by different scorers?
   b. What are the principles for placing anchor samples in stacks, especially for longitudinal analysis of the application of the standard over time?
4. Analysis of the raw data from scoring collected in moderation. Decisions related to the suitability of item response models (e.g., the Rasch model) to provide evidence of the application of the standard. This evidence is used to determine if TPAs are assessing graduates at a comparable level and the reliability of judgment. Judge severity and lenience is another consideration.

Phase 3 extended to issues of sustainability of the moderation cycle and associated activities, and the potential for growth. A further consideration was the potential of including a larger group of universities, including those internationally. Specifically, goals were identified as a need to:

1. Support the GTPA Collective with customized apps to enable data collection, collation, storage, and visualization. These apps facilitated access to an integrated database that used the latest security features and was easy to use by non-specialists in digital data management.
2. Enable reporting on patterns at the level of overall performance assessed against a standard and criterion-level scores for program cohorts.
3. Undertake longitudinal monitoring of trends in the annual ITE program data and examine comparability in applying the standard across universities and within and across programs in a single institution.

The use of digital technologies and a customized infrastructure was imperative to respond to these goals. The infrastructure that was designed to address these goals included a software system and a data warehouse. We named the system *Evidence for Quality in Initial Teacher Education* (EQuITE). The development of EQuITE, as a central source for evidence collection and analysis, is significant in enabling outcomes for improving the quality and impact of teacher education. It includes a means of translating the results from complex statistical analyses of the data into a form that is recognizable to teacher educators using data visualization.

The EQuITE software system and data warehouse has five customized components:

1. Online submission of selected GTPA performance samples for the purpose of CIM. This includes storage of GTPA sample files and the corresponding database containing details of the sample. Scores collected from CIM are recorded in the data warehouse against the sample identifier and contextual data about the sample.
2. Web portal for online cross-institution scoring of samples against the established standard (GTPA CIM-Online™). These processes are detailed below.

3. The GTPA App for online submission, collation, and storage of program demographic and cohort performance data. The app was developed to address the considerable barriers to collating program cohort data for and from multiple institutions while ensuring security and privacy of data. Universities action the submission of their chosen data which is automatically deidentified at the point of transfer.
4. Data analysis and automated report generation to provide confidential evidence to universities from their own data.
5. Access to a repository of deidentified longitudinal data on consistency of judgment, application of the standard, and patterns of ITE program cohort scores, contributed by the GTPA Collective. This enables large-scale data collection over time to provide evidence on the standard of preservice teacher preparedness for classroom teaching in Australia as well as feedback to teacher educators for ITE program improvement where this is needed.

Benchmarking for external verification of the application of the standard across sites

In this section, we describe what teacher educators do when participating in GTPA benchmarking through CIM-Online™. Teacher educators from each participating university self-identify to participate as raters. They work independently to rescore samples chosen by universities to represent the full range of quality. The rater's role is to determine if the sample meets the standard at or above the minimum accepted level. Our interest is in benchmarking within a single TPA, that is the GTPA, recognized as the largest TPA collective of universities in Australia.

In GTPA CIM-Online™, all teacher educator raters are experienced. The approach to selecting raters is to ensure coverage of a range of content areas, all phases of schooling (early years, middle years, upper secondary), and diversity of expertise across degree programs. The scoring of samples is undertaken by raters through an online web portal. CIM-Online™ has the advantage of enabling raters to participate from their own locations.

The design process for undertaking benchmarking begins with addressing several considerations. These include: (1) the total number of samples to be scored, (2) the distribution of received samples across the quality range, (3) the number of scores required to measure consistency in scoring for a sample, (4) the number of raters available for scoring, and (5) the number of samples that a rater could be expected to score given their agreed time allocation.

Raters are allocated a selection of de-identified GTPA samples presented as virtual stacks. The samples have been scored and moderated previously within each participating university. The samples are chosen from across the quality range. Raters do not score their own university samples. Allocation of GTPA samples to raters is additionally based on the principle that performances of quality close to the threshold (low pass at the minimum acceptable level) are scored by at least ten raters. Samples considered by the submitting university to be clearly above or below the threshold are scored by three to six raters. Rater workload is contained to an acceptable level, as agreed by the Collective.

A related priority is generating sufficient scores to compute an overall measure of reliability with a reasonable degree of accuracy. An outcome of CIM is to demonstrate the reliability of judgment against an established standard. Recognizing that standards can rise and fall over time, anchor samples with previously endorsed judgments are embedded in virtual stacks for the new scoring round. Their utility is the application of the standard applied in previous rounds. In this way, they enable monitoring movement in the standard over time. To illustrate the process, in the 2021 scoring activity, 118 raters scored 253 samples including two anchor samples that had been included in previous CIM-Online™ activities.

The inclusion of anchor samples in the moderation event ensures that the opportunity for monitoring the movement of the standard over time is not compromised. Without this monitoring, it could be argued that it was easier to graduate from teacher education last year or the year before. This raises issues of fairness (as discussed in Chapter 5). Legal precedence for cases contesting grading decisions in the case of the edTPA and PACT are instructive for Australia (see Chapter 10).

Scoring

The audit of the GTPA, discussed in Chapter 4, brought to light the nature of the criteria as they aligned to the Australian Professional Standards for Teachers (APST; Australian Institute for Teaching and School Leadership [AITSL], 2011). While the APST have served as inputs into program design and accreditation, they had not been associated with the requirement for the moderation of judgments of a culminating performance assessment in teacher education. A design feature of the GTPA was that the criteria functioned to provide multiple opportunities for preservice teachers to demonstrate identified professional standards or APST. The following discussion addresses the function of the criteria in grading decisions at two levels. The first concerns the advice provided to raters about the grading process. Raters were advised to:

1. Familiarize themselves with the criteria prior to reading the sample.
2. Read the sample and arrive at an initial interim assessment (Meets/Does Not Meet).
3. Apply an analytic approach to each of the practices in turn against the stated features in the criteria sheet. Record a score against each criterion.
4. Finalize the grading decision when the interim assessment and the outcome of the analytic approach are consistent.
5. Review the judgment to identify the influence of unstated features when the interim assessment and the outcome of the analytic approach are inconsistent.

The second level concerns what was learned about the function of criteria in actual grading decisions. The analysis of the grading process at the criterion level has brought to light new information about patterns of performance across the criteria and the influence of trade-offs or compensations. Specifically, the analysis has revealed those aspects of performance that are stronger, those that are weaker, and how they combine in overall judgment. Trade-offs remain invisible – they leave no trace of their influence in judgment in the final score as recorded on a script.

We did not assume that the criteria alone wholly regulate individual judgment or necessarily assure reliability. The analytic methodology applied in the GTPA study showed the severity and leniency of individual raters relative to the pool of scorers involved. Severity and lenience typically remain invisible in the judgment process, though some judges take it as a badge of pride to say that they are 'hard markers'. See Chapter 8 for the analysis of judge severity from GTPA CIM-Online™ activities and how these are used in reports provided to universities.

Analysis of data from scoring performances

Analysis of data from scored performances is undertaken by applying a multi-facet Rasch model (MFRM; see Chapter 6) to the quality of performance of a sample relative to three underlying traits: (1) performance ability of the preservice teacher completing the GTPA sample, (2) difficulty levels of the five criteria of assessment, and (3) the severity of raters in making a judgment about the sample. The performances of preservice teachers are measured by their estimated location on the common logit scale and ranked in order from lowest to highest. The scoring data, from CIM and ITE program performance data collected through the GTPA App, informs the production of an automated report with accompanying visualization of the analyzed data. Post-analysis moderation meetings, which are conducted online or in-person with each ITE provider, involve discussion of the moderated grades (see Chapter 8).

Step 3: Designing to build dependability: Fidelity of implementation, material artifacts, scorer training, and calibration

When TPAs were introduced into Australian teacher education, the standard representing profession readiness had not been established and large-scale CIM had not been a feature in the culture of teacher education. Moderation of the GTPA involves the application of an established, agreed standard of graduate readiness for professional practice by teacher educators as experts with professional guild knowledge. The generation of the standard and its application were new practices for teacher educators (see Chapter 6 for discussion of setting the standard).

While digital architecture is essential to undertake CIM and the collection of data at scale in ITE reform, the enablers for this system are still talk, text, and interaction. Purposefully designed in-person and online events were established as a core feature of the Collective's activities to encourage the sharing and build of knowledge, skills, and understanding. These discussions centered around the use of artifacts designed to support the implementation, judgment, and moderation activities.

Training is an ongoing feature of the GTPA Collective meetings where validated GTPA samples are discussed to understand expected characteristics of a level (e.g., Meets at the minimum). In addition, underpinning the practices associated with implementing this new complex performance assessment were principles of fidelity to ensure the dependability of the assessment.

Following is a description of the essential preparatory work for undertaking benchmarking including (1) establishing principles of fidelity, (2) designing the resources – decision aids – used to inform judgment, and (3) implementing calibration activities to promote the reliability of judgment.

Fidelity of implementation

The term fidelity has been used in education in relation to the degree to which simulations represent real-life scenarios (Caliço, 2017), and in medicine, in terms of adherence to a treatment protocol. In educational assessment theory, Sadler (2010) positions fidelity as "a precondition for integrity in grading academic achievement" (p. 727). Thus, the grade is determined only on, and is true to, the elements that relate to what is being assessed. It is not determined based on factors or 'contaminants' such as effort and attendance. A grade given at the completion of a unit of work "should represent learners' attained levels of academic achievement" (Sadler, 2010, p. 728). Fidelity in assessment is a key consideration since "any lack of fidelity places an upper bound on the maximum achievable level of validity" (Sadler, 2010, p. 729) and is a necessary condition and a precondition for establishing reliable evidence about achievement when using established standards for grading.

During the Phase 1 stage of GTPA validation and standard setting (see Figure 7.1, Table 7.1), it was evident that there were divergent procedures and practices across the country in assessing the achievement of teacher education students, both in their university academic program and in their professional experience placements in schools. The reach of the GTPA across Australian states and territories demanded attention to the conditions under which the GTPA was to be implemented and the consequences of these conditions for valid and reliable assessment. Further, we took the position that "without specific, clear measurement of implementation, it is impossible to know whether disappointing outcomes are due to an inadequate program model of change or due to poor or incomplete implementation" (Century et al., 2010, p. 200). Consideration was given to how closely the design and intent of the assessment were followed in practice, and how the overall structure of the program, that is the sequencing and development of learning, sets students up for success.

At this stage, our investigations turned to aspects of GTPA implementation that needed to be consistent across all ITE programs to maintain system validity, and those aspects that could be varied to accommodate the site-specific characteristics of ITE programs, school practicum sites, and student cohorts (site validity; see Chapter 3). Fidelity of implementation of the GTPA was deemed a necessary precondition for building stakeholder confidence in decisions regarding preservice teacher

TABLE 7.2 Examples of personal, program, and system considerations that may influence the fidelity of implementation

Personal	*Program*	*System*
• Teacher educators' understanding of the GTPA • Professional development to support GTPA delivery • Degree and timing of formative assessment provided to the preservice teacher • Quality, accessibility, and use of resources by teacher educators and preservice teachers • Access to data and evidence provided to preservice teachers during school placement • Teaching conditions and support provided to preservice teachers during school placement	• Alignment of the assessment with the Graduate Teacher Standards (AITSL, 2011) • Progressive development of knowledge and skills across the ITE program • Timing and duration of the final-year school placement • Moderation processes and practices to support comparability of judgments	• University policy and procedures for aspects such as course design, course changes, assessment timing, and weightings • Teacher education regulatory authorities' policies for final school placements • Teacher union policies including the workload of supervising teachers and teacher educators

preparedness for professional practice (Adie & Wyatt-Smith, 2018). Questions were identified for consideration of how and under what conditions the GTPA was being implemented. These included: What are teacher educators' understanding of the core tenets underpinning the assessment design and the critical components of the assessment? What should the alignment of a program, that will build expertise and support completion of the assessment, look like? What aspects of the assessment design and implementation can be adapted to local contexts while still maintaining the fidelity of the assessment? What practices will result in the fidelity of the assessment being compromised? Table 7.2 provides an example of the aspects that were considered in our investigation of the conditions of GTPA implementation, organized in three levels: personal, program, and system.

Consideration of the aspects identified in Table 7.2 led to the establishment of principles of fidelity. The principles needed to maintain the integrity of the assessment according to the standards and procedures for the accreditation of ITE programs in Australia (AITSL, 2015; see Chapter 4), and ensure equity of opportunity for preservice teachers completing the GTPA.

For the GTPA to be perceived as dependable, implementation conditions should be recognized as 'like' or comparable. For example, preservice teachers require knowledge of the Graduate Teacher Standards (AITSL, 2011) and teacher educators and preservice teachers need to recognize how the standards are infused into their learning, both in the academic program and during school placements. In their final year school placement, in which the GTPA is completed, it is a reasonable expectation that there are opportunities to demonstrate professional competence and to collect the necessary evidence to be assessed as meeting the requirements of the Graduate Teacher Standards (AITSL, 2011).

Equity of opportunity to complete the GTPA also requires that the assessment remains intact, completed as a whole, without any additions or changes, and implemented, completed, and submitted within established and agreed conditions and timeframes. It should not be submitted in separated segments or components over time. Preceding professional experience assessments and academic program assessments need to be understood as preparation for the GTPA and not as evidence contributing to its completion. For example, a cumulative assessment such as a portfolio compiled over the period of the program is not consistent with the intent of the GTPA to be a culminating assessment of competence completed in the final year of an ITE program. In addition, evidence collected from a different school or a different class at an earlier period is also not valid evidence of teaching as a connected activity of *Planning*, *Teaching*, *Assessing*, *Reflecting*, and *Appraising*. The GTPA is purposefully designed to demonstrate the iterative nature of teaching in which teaching plans are developed and then adjusted as information is continually received from students during an uninterrupted teaching episode. The evidence collected early in the school placement is to be revisited over time and throughout the teaching to update understandings about student learning and to inform instructional decision-making.

Other considerations in establishing principles for fidelity included the distinction between purposes of formative and summative assessments. The official function of TPAs was as a high-stakes, summative assessment completed in the final year of the program. Early in the implementation of the GTPA, a threat to fidelity was teacher educators' use of formative feedback in supporting preservice teachers to complete the assessment. A completed GTPA leaves no trace of the nature and extent of direct input from others, confounding its utility to represent preservice teachers' competence. Validity and fairness issues surrounded the possibility that it was teacher educators' knowledge dominant in the submission (and not that of preservice teachers). Torrance (2007) referred to the limited interpretation of formative assessment in higher education as "an overwhelming focus on criteria compliance and award achievement" (p. 282). The result is preservice teachers who are dependent on the teacher educator's advice rather than themselves as self-regulated learners – a result which is contrary to that of the GTPA as an authentic representation of preservice teacher readiness for the profession. Formative assessment as used with the GTPA needs to encourage a preservice teacher's reflective, reasoning, and problem-solving practices rather than provide them with an answer or nudge them toward the correct answer (for further discussion of this point, see Gelfuso, 2017; Rodgers, 2006; Torrance, 2007).

In establishing the principles, we recognized that authentic assessments such as the GTPA occur in localized contexts which calls for system validity to be balanced against site validity. The principles of fidelity needed to be inclusive of the different school and community contexts in which preservice teachers may complete their GTPAs as well as ensure that the integrity of the awarded grade is a true representation of "the quality, breadth and depth" (Sadler, 2010, p. 728) of the preservice teacher's knowledge and skills. The assessment should produce grades that are trustworthy, representative of quality, and are a true indicator of achievement for the student, the institution, the employer, and for future clients (e.g., students). When establishing principles of fidelity, we were ensuring a level of stability across samples (system validity) such that samples are scored according to the specified scoring rubric and by teacher educators who have relevant discipline and pedagogic expertise, while maintaining flexibility for site variables (site validity). These conditions are necessary for rigorous moderation processes in which teacher educators maintain focus on system requirements while accommodating for site or contextual variabilities.

Principles of fidelity shine the light on the conditions under which the GTPA is to be implemented. Working toward fidelity of the assessment, several procedures and related resources have been developed. These include guides and associated resources for preservice teachers completing the GTPA; fact sheets targeted at different stakeholders to ensure an informed community; and organized online meetings for teacher educators and regulatory authorities to discuss insights and issues and engage in collaborative problem-solving. Of significance to CIM are the decision aids used to promote a shared understanding of the stated standard, described next.

Decision aids

A suite of decision aids that clarifies the expected standard and associated criteria has been developed to support judgment and moderation activities. The decision aids include performance level descriptors (PLDs), criteria specifications, exemplars selected from validated GTPAs, and cognitive commentaries (Figure 7.2). These have been informed by the APST (AITSL, 2011) and Program Standards and Procedures (AITSL, 2015).

1. *PLDs* describe graduate performance characteristics at four levels – Meets, Above, Below, and Minimum accepted performance. The descriptions at each level identify anticipated features of performance aligned with the criteria but are not expected to wholly define performance. They refine the gaze of judges to the different patterns evident at levels of performance. The primary purpose of the PLDs is to support judgments of the quality of a performance. The PLDs are used to train raters to identify anticipated features at, above, and below the standard, and the different ways these features can be demonstrated across samples. See Chapter 6 for the PLDs used in Phase 1 Validation and Standard Setting, noting that these have been refined in response to the trial.
2. *Criteria specifications* identify the APST that are evident in each of the five GTPA practices – *Planning, Teaching*, *Assessing*, *Reflecting*, and *Appraising* (see Chapter 4). The assessment is designed so that there are multiple opportunities for demonstrating the APST across each practice.

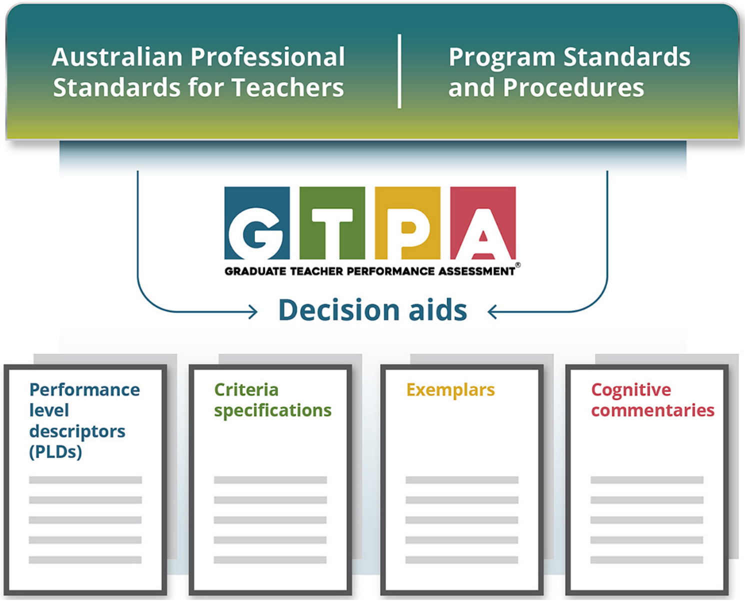

Australian Professional Standards for Teachers
Program Standards and Procedures
GTPA
GRADUATE TEACHER PERFORMANCE ASSESSMENT®
Decision aids
Performance level descriptors (PLDs)
Criteria specifications
Exemplars
Cognitive commentaries

FIGURE 7.2 The decision aids that support the GTPA moderation processes

3. *Exemplars* taken from validated GTPAs illustrate different levels of performance. This is the next layer to refine the gaze of judges as they interrogate different ways to demonstrate a level of performance, as well as differences across levels.
4. *Cognitive commentaries* are the written explanation of how a judgment was made, considering the strengths and weaknesses of a performance. The commentaries explicate or bring to the surface trade-offs that would otherwise remain unstated by the assessor and as such are invisible aspects of judgment. Raters use the commentaries to understand the intricacies of a judgment decision, understanding that different features may come to the fore across samples judged to be at the same standard of performance.

The decision aids have been designed to reach into the depths of judgment decisions. Their combined use addresses the purpose of making clear the meaning of the criteria and standard. Engagement with the decision aids develops an understanding of the task and the related standard and criteria with the aim to increase the reliability of judgments, and thus dependability.

Calibration

The term calibration refers to training in how to arrive at a scoring decision. The aim is to align grading decisions to the established standard of Meets such that all raters recognize similar qualities as representative of the standard. The purpose of calibration is to build capability in making judgments according to the established standard, that is, to promote the reliability of judgments. In the GTPA Collective, calibration activities are scheduled through online meetings throughout the year as well as through a designated event that is conducted online prior to commencing CIM.

Calibration training is a critical step in preparation for CIM scoring. All GTPA raters must undergo calibration before commencing their blind re-scoring of submitted samples. This requirement is grounded in the research finding that standards applied by even experienced assessors do not necessarily remain stable (Bloxham, 2009). Each rater scores three samples which have been verified through a previous GTPA CIM-Online™ event.

Raters identify each sample as Meets, Above, or Does Not Meet the standard. Upon submission of their judgments, they are supplied with the verified level (Meets, Above, or Does Not Meet) and the associated cognitive commentary for the sample. Cognitive commentaries are written specifically for each sample using the criteria specifications and the PLDs as a basis for the descriptions of performance. That is, the cognitive commentary is written using the specific features of the performance rather than the general level of description in the criteria or descriptors.

Raters complete calibration before commencing scoring and are able to revisit the calibration samples for review during scoring as appropriate. In this way, raters are being trained in the application of the criteria and what this may look like in actual performances. Calibration activities prior to CIM use an online system for the immediate return of results from calibration training. Where necessary, raters receive a second set of samples for further calibration. In this case, the cognitive commentary is provided with the sample, so that raters are trained to see critical evidence identifying a level of performance.

The importance of calibration training in promoting judgment reliability is clear. Over the period 2018–2021 in the GTPA Collective, there has been an increasing rate of endorsement of scores through CIM as participation in calibration increases (Figure 7.3). To examine the impact of this increase on rater reliability, we analyzed internally moderated results provided by individual universities with externally moderated scores produced from CIM. Scores were endorsed if there was agreement between scores from internal and external moderation. Such increases are a positive sign indicating the potential of calibration in enabling the consistent application of a profession-ready standard across universities. Public confidence in the consistent application of an agreed and established standard depends on such evidence. This has not been available previously. Further, the data provide evidence of the benefit of a sustained focus on calibration training to build reliability.

Conclusion

In the GTPA Collective, we have taken the firm position that CIM is an essential part of quality and assurance systems and processes necessary to lift the status of the profession and improve confidence in teacher preparation. Our work has extended to innovation in analysis and reporting, and the use of digital infrastructure for universities to demonstrate reliability. The chapter has introduced a new conceptualization

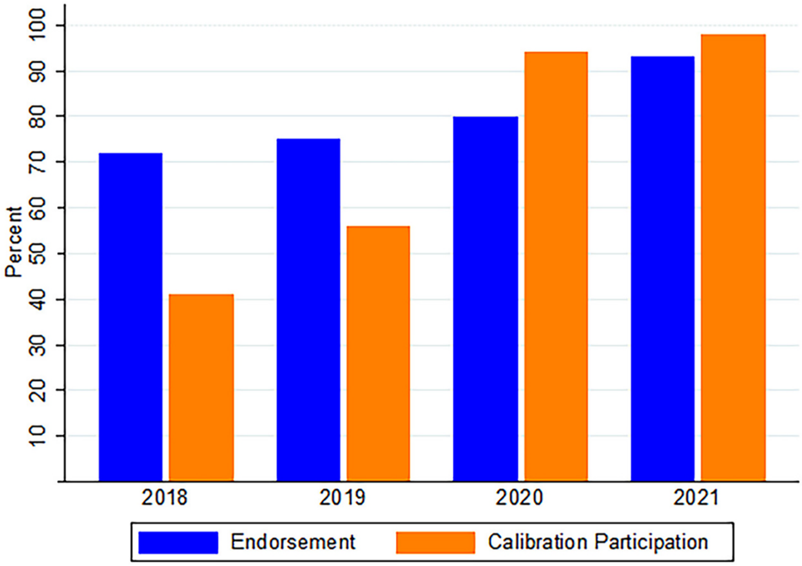


FIGURE 7.3 Endorsement rate and calibration participation over time (2018–2021)

of moderation that carries forward elements of both social and statistical moderation, and shifts moderation into an online space involving human–machine interactions. Digital applications have been critical for sustaining the practice and going to scale.

In the introduction to this chapter, we proposed five steps that we took in our approach to GTPA CIM-Online™. This chapter has addressed the first three steps. The next chapter will address the final two steps that show the use of analyzed data from CIM in the production of customized reports for each participating university. We introduce a new feedback loop in ITE through which summative data can be used for curriculum review and program renewal purposes. The use of this type of data is new in teacher education and requires teacher educators to learn how to interpret reports and infer meaning from them. Moderation is positioned at the heart of efforts for agentic action by teacher educators, enabling them to use data to present voices in policy-driven reform. The collective efforts have moved well beyond policy-driven compliance. They have fostered and installed an inquiry approach in teacher education that has at its core the goal of improving learning for all school students.

Notes

[1] Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

2 Acknowledgment: The online model of cross-institutional moderation (CIM-Online™) was conceptualized and developed in the Institute for Learning Sciences and Teacher Education, Australian Catholic University. For a discussion of CIM-Online™, readers are advised to also see Wyatt-Smith and Adie (2021).

References

Adie, L. (2013). The development of teacher assessment identity through participation in online moderation. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 91–106. https://doi.org/10.1080/0969594X.2011.650150

Adie, L., & Wyatt-Smith, C. (2018). Research-informed conceptualization and design principles of teacher performance assessments: Wrestling with system and site validity. In C. Wyatt-Smith & L. Adie (Eds.), *Innovation and accountability in teacher education: Setting directions for new cultures in teacher education* (pp. 115–132). Springer.

Adie, L., & Wyatt-Smith, C. (2020). Fidelity of summative performance assessment in initial teacher education: The intersection of standardisation and authenticity. *Asia-Pacific Journal of Teacher Education*, *48*(3), 267–286. https://doi.org/10.1080/1359866X.2019.1606892

Australian Institute for Teaching and School Leadership (AITSL). (2011; revised 2018). Australian professional standards for teachers. https://www.aitsl.edu.au/teach/standards

Australian Institute for Teaching and School Leadership (AITSL). (2015; revised 2018, 2019). Accreditation of initial teacher education programs in Australia. https://www.aitsl.edu.au/docs/default-source/national-policy-framework/accreditation-of-initial-teacher-education-programs-in-australia.pdf?sfvrsn=e87cff3c_28

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment and Evaluation in Higher Education*, *34*(2), 209–220. https://doi.org/10.1080/02602930801955978

Bloxham, S., Hughes, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment and Evaluation in Higher Education*, *41*(4), 638–653. https://doi.org/10.1080/02602938.2015.1039932

Caliço, T. (2017). Graduate student corner: Being an assessment professional in games-based assessment. *National Council on Measurement in Education Newsletter*, *24*(4), 5–7.

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *32*(2), 199–218. https://doi.org/10.1177/1098214010366173

Crimmins, G., Nash, G., Oprescu, F., Alla, K., Brock, G., Hickson-Jamieson, B., & Noakes, C. (2016). Can a systematic assessment moderation process assure the quality and integrity of assessment practice while supporting the professional development of casual academics? *Assessment and Evaluation in Higher Education*, *41*(3), 427–441. https://doi.org/10.1080/02602938.2015.1017754

Crisp, V. (2017). The judgement processes involved in the moderation of teacher-assessed projects. *Oxford Review of Education*, *43*(1), 19–37. https://doi.org/10.1080/03054985.2016.1232245

Estyn. (2016). *Moderation of teacher assessment at key stage 2 and key stage 3: A review of accuracy and consistency*. Cardiff, Wales. https://www.estyn.gov.wales/sites/www.estyn.gov.wales/files/documents/Teacher%20assessment%20-%20Eng.pdf

Gelfuso, A. (2017). Facilitating the development of preservice teachers' Pedagogical Content Knowledge of literacy and agentic identities: Examining a teacher educator's intentional language choices during video-mediated reflection. *Teaching and Teacher Education*, *66*, 33–46. https://doi.org/10.1016/j.tate.2017.03.012

Gilmore, G., Margrain, V., & Mellgren, E. (2020). Intercultural literacy dialogue: International assessment moderation in early childhood teacher education. *Intercultural Education*, *31*(2), 208–227. https://doi.org/10.1080/14675986.2019.1702293

Harlen, W. (2004, 29 November). Can assessment by teachers be a dependable option for summative purposes? General Teaching Council for England Conference, New Relationships: Teaching, Learning and Accountability, London. In C. Adams & K. Baker (Eds.), *Perspectives on pupil assessment* (pp. 24–30). https://dera.ioe.ac.uk/14022/1/1104_Perspectives_on_Pupil_Assessment._New_Relationships__Teaching,_Learning_and_Accountability.pdf

Hipkins, R., & Robertson, S. (2012). The complexities of moderating student writing in a community of practice. *Assessment Matters*, *4*(2012), 30–52.

Klenowski, V., & Wyatt-Smith, C. (2010). Standards-driven reform years 1–10: Moderation an optional extra? *Australian Educational Researcher*, *37*(2), 21–39. https://doi.org/10.1007/BF03216920

Pobble (n.d.). *The future of writing moderation*. https://my.pobble.com/moderation

Rajamany, V., van Biljon, J., & van Staden, C. (2020). *eModeration adoption requirements for secondary school education: A critical literature review*. Conference on Information Communications Technology and Society (ICTAS), https://doi.org/10.1109/ICTAS47918.2020.233979

Rodgers, C. R. (2006). Attending to student voice: The impact of descriptive feedback on learning and teaching. *Curriculum Inquiry*, *36*(2), 209–237. https://doi.org/10.1111/j.1467-873X.2006.00353.x

Sadler, D. R. (2010). Fidelity as a precondition for integrity in grading academic achievement. *Assessment and Evaluation in Higher Education*, *35*(6), 727–743. https://doi.org/10.1080/02602930902977756

Tarricone, P., & Newhouse, C. P. (2016). A study of the use of pairwise comparison in the context of social online moderation. *Australian Educational Researcher*, *43*(3), 273–288. https://doi.org/10.1007/s13384-015-0194-z

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education: Principles, Policy & Practice*, *14*(3), 281–294. https://doi.org/10.1080/09695940701591867

Wiliam, D. (1994). Reconceptualising validity, dependability and reliability for National Curriculum assessment. In D. Hutchison & I. Schagen (Eds.), *How reliable is national curriculum assessment?* (pp. 11 – 34). National Foundation for Educational Research. https://www.nfer.ac.uk/media/1422/91098.pdf

Williamson, J. (2016). Statistical moderation of school-based assessment in GCSEs. *Research Matters: A Cambridge Assessment publication*, *22*, 30–36. http://www.cambridgeassessment.org.uk/research-matters/

Wyatt-Smith, C., & Adie, L. (2021). Introducing a new model for online cross-institutional moderation. In C. Wyatt-Smith, L. Adie, & J. Nuttall (Eds.), *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration* (pp. 39–58). Springer.

Wyatt-Smith, C., & Adie, L. (in press). The role of teachers in making and moderating assessment judgements: Has the black box been opened in Australia? In C. Harrison, C. Leung & D. Pepper (Eds.), *A festschrift for Paul Black*. Bloomsbury.

Wyatt-Smith, C., & Colbert, P. (2014). An account of the inner workings of standards, judgement and moderation: A previously untold evidence-based narrative. Informing paper for the Review of Queensland Senior Assessment and School Reporting and Tertiary Entrance Processes undertaken by Australian Council for Educational Research (ACER). Brisbane, Australia: Learning Sciences Institute Australia, Australian Catholic University. http://www.acer.edu.au/files/WyattSmithColbert_InformingPaper_Final.pdf

8

WHY TEACHER EDUCATION NEEDS A FEEDBACK LOOP

Connecting standards and evidence to inform program planning and renewal

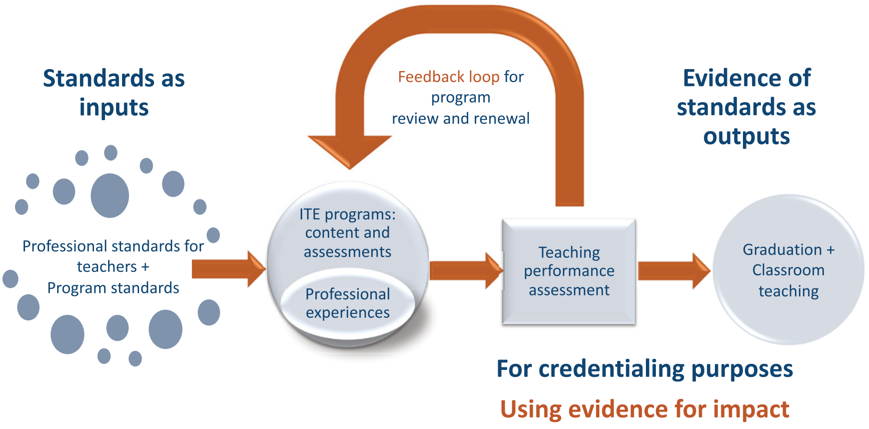
Introduction

This chapter presents a new configuring of the purposes of teaching performance assessments (TPAs). We discuss the potential of analyzed TPA data to strengthen the teacher education workforce, and in turn, the teaching workforce. Our focus is on three main concepts: performance data, reporting, and use of the data by teacher educators, regulatory authorities, preservice teachers, and the wider community. These come together in the core notion of considering fitness-for-purpose. In our experience, Australian teacher educators have been eager to have customized data that they can use as an evidence base for summative reporting of TPA performance. Beyond this, they have been eager to receive reports of data with formative potential. These have initiated new types of conversations that root research, practice, and policy in local data and the larger corpus of data drawn from the group of universities participating in the Graduate Teacher Performance Assessment (GTPA®).[1]

We explore the purposes and value of GTPA outcomes reports produced from cross-institutional moderation (CIM-Online™[2]; see Chapter 7). This type of data has not been available previously. We identify the need for teacher educators to develop data literacies for interpreting and using evidence to inform program improvements. Our starting proposition is that validated TPAs are necessary, but insufficient, in and of themselves to improve teacher education. The next step in the maturation of TPAs is how to connect the summative (reporting) purposes of the data and their uses for formative (improvement) purposes. When this connection is made, credentialing purposes and program review purposes can be brought into scope. The data can be customized to establish a feedback loop from program outputs to program inputs (see Figure 8.1).

Our use of the term 'feedback loop' links the formative purpose of assessment with standards and evidence. Ramaprasad (1983), working in the field of management

DOI: 10.4324/9780429318504-10


Standards as inputs
Feedback loop for program review and renewal
Evidence of standards as outputs
Professional standards for teachers + Program standards
ITE programs: content and assessments
Professional experiences
Teaching performance assessment
Graduation + Classroom teaching
For credentialing purposes
Using evidence for impact

© Australian Catholic University Limited (ABN 15 050 192 660)

FIGURE 8.1 A feedback loop for teacher education connecting standards as inputs to standards as evidence of outputs

theory, described feedback as information used to alter "the gap between the actual level and the reference level of a system parameter" (p. 4); it is only when information is acted upon to alter the gap that the feedback loop is complete. Sadler (1989), applying Ramaprasad's (1983) work to the field of education, added the necessity for both teachers and students to have the knowledge and skills to identify quality performance and act on feedback to improve performance. The intent is that over time, dependence on the teacher as the sole source of feedback lessens as students' evaluative expertise develops. In the GTPA project, feedback is provided via the analyzed CIM data and customized reports to each participating university. Teacher educators who receive these reports need to have developed knowledge of the required standard of performance as well as the skills to interpret the data and inquire into, and improve, program effectiveness. Knowledge of the standard is developed through calibration activities that occur individually and online using exemplars and cognitive commentaries (explanations of judgment decisions) and through discussions of judgment decisions during in-person and online Collective meetings (see Chapter 7 for a description of calibration activities including the use of cognitive commentaries). Teacher educators interpret the CIM report by drawing on their shared understanding of the established standard used to judge GTPA performances and their knowledge of program and professional standards. By linking these related fields of knowledge, teacher educators can inquire into the quality of their programs and make decisions about improvement actions. It is in these actions that the loop, from TPA outcomes to curriculum review and program renewal, is formed.

Historically, professional and program standards have been used to inform program design and accreditation purposes (see Chapters 1 and 4 for a discussion of standards). Our interest in the discussion that follows is how evidence of standards met through a TPA can function as a feedback loop.

In what follows, we first provide two perspectives on the functions of standards in teacher education (1) as restrictive and (2) as enabling. Next, we describe evaluative expertise and data literacy as necessary skills in the productive use of TPA data. We describe how GTPA data can contribute to a cycle of program improvement, supported through digital infrastructure. We illustrate aspects of the customized reports that participating universities receive. These provide two categories of data analysis. The first presents outcomes from analyzing CIM-Online™ scores and shows the application of the standard in specific programs. The second uses raw data – self-reported by each university – in the form of cohort performance information at the criterion level. Finally, we assert the value of a feedback loop informed by data from multiple institutions to show how a common standard has been applied.

Standards as restricting and enabling

Professional standards are a recognized hallmark of a profession. In teacher education, standards have functioned as program referents: evidence is required of how they have been introduced, developed, and assessed across teacher education programs. This use of standards has received much criticism. Professional standards have been described as "boss texts" that are part of a "managerial agenda" that seeks to govern, shape, and regulate teachers' work and learning (Talbot, 2016, p. 81). While standards have received praise as providing transparency of expectations and a necessary component of a highly skilled workforce (Darling-Hammond, 2012), they have also been criticized as technicist, responsive to neoliberal accountability agendas (Connell, 2009; Delandshere & Arens, 2001; Talbot, 2016), and leading to a checklist of attributes that restrict deep learning and overlook situated circumstances (Connell, 2009).

Critics of professional standards point to a mistrust in the professionalism of teachers. For example, Connell (2009) discussed the intellectual work of teachers as they analyze and adapt to each student response in the complexity of classroom interactions. The codification of this work, Connell claimed, promoted a tick-box approach to teacher education and allowed professional "control at a distance" (p. 222). Following the introduction of professional standards in two states in the United States, Delandshere and Arens (2001) lamented that "teacher educators are no longer autonomous intellectual agents as they are constantly wondering what they will have to implement next. When teaching becomes the implementation of limited vested political ideas it loses any social, political and intellectual ideal or engagement" (p. 564). In this context, TPAs, as informed by standards, are open to criticism of "contributing to existing structures of power and repression" (Talbot, 2016, p. 83). This occurs especially when TPAs are introduced as part of a top–down reform strategy.

Shifting the focus on standards-as-inputs to standards-as-outputs has also been criticized as invoking "factory and assembly-line images of schooling" that "assumes a linear relationship between teaching and learning for both K-12 students and for teacher candidates" (Cochran-Smith, 2004, p. 205). Standards can function as checklists of separate items, seemingly unrelated, able to be ticked off when completed. In these narrow uses, standards may not reflect the complexity of actual performance, obscuring the nuances of contextually relevant, situated responses. The restrictive potential of standards on the work of teacher educators led Delandshere and Arens (2001) to suggest that "teacher education institutions should be engaged in defining and studying cases of their teaching and in providing evidence of their work" (p. 564).

We present a conceptualization of standards that has a dual interest in measurement and teacher agency. Our approach is not an atomistic or mechanistic use of standards to assess competence. Rather, we suggest a focus on the professional judgment and evaluative expertise required of teacher educators as they interpret and use standards in a range of ways to evidence authentic professional practice that is responsive to context. We lend support to Darling-Hammond's (2008) observation that standards, in and of themselves, will not improve teacher education and, dependent on how they are used, could be restrictive. Our position is that when standards are used with a focus on evidence to show graduate competence as well as the quality and impact of ITE programs, teacher educators can exercise professional accountability. Such a system involves front-ending standards for program planning and review, going back to standards as (1) quality assuring graduate readiness and (2) showing the impact of teacher education programs on graduate learning.

The term 'front-ending standards' for program planning refers to the use of standards to inform the design of teacher education programs in contextually relevant ways (Wyatt-Smith & Bridges, 2008). The notion of 'front-ending' is situated in a sociocultural understanding of teacher education as responsive to both system and local community contexts (see Chapter 3). In employing the standards, teacher educators draw on their understanding of the communities they serve, to interpret the standards and design programs that incrementally develop generic skills responsive to community needs.

The understanding of standards in the design of the GTPA and related scoring criteria has been grounded in the potential of standards to have both formative and summative purposes such that the generated evidence will inform curriculum review and program renewal. The feedback loop occurs when data – moderated teacher judgments – is fed back to the participating universities via extensive reporting of results. The reports, customized for each university, provide evidence of scoring consistency against an established common standard of achievement, and the strengths and weaknesses of ITE programs relative to the empirically derived GTPA standard (see Chapter 6).

The evidence in the annual GTPA Report is interpreted and used by teacher educators working in their local contexts to inform conversations about program quality and identified areas of underpreparation. The focus here is on preservice teachers' learning and how the ITE program supports professional preparation. This process moves beyond a direct linking of standards at the beginning of program design to the development of graduate capabilities in the course of the program. Rather than standards being used as a tick-box of completion that produces a simplified codification of the complexity of teacher education, the standards can be used "as reflective and planning tools exploring their contextualized practice" (Forde et al., 2016, p. 31). Vital in this process is the generation of data from the final competence assessment, now fed into a feedback loop that connects assessment, learning, and teaching. The feedback loop is key to teacher educators maintaining agency within this work or, as Darling-Hammond (2012) described, to "take charge of accountability and make it useful for learning and improvement" (p. 13).

Standards are the informing structure for the feedback loop and the generation of evidence. The use of evidence from graded GTPAs has feedforward potential when used by teacher educators. This distinguishes the use of evidence in the GTPA from claims that have been made about TPAs and program development elsewhere. For example, the edTPA, in use across approximately 700 U.S. universities, also claims to "guide the development of curriculum and practice around the common goal of making sure new teachers are able to teach each student effectively and improve student achievement" (Stanford Center for Assessment, Learning & Equity, n.d., n.p.). However, evidence indicates that this guidance refers to how the assessment and specifically, how it is designed, can influence program design (Cohen et al., 2018). Table 8.1 summarizes the relationship between the standards and the GTPA at the input and output points in the feedback loop.

In the GTPA project, our efforts are focused on connecting standards as part of a formative assessment strategy in which feedback leads to an investigation of practice. The potential of quality feedback to improve practice or performance is well documented (Black & Wiliam, 1998; Hattie & Timperley, 2007). Feedback can help to identify gaps in knowledge and thus areas requiring further attention. Hattie and Timperley (2007) identified three core questions that feedback can address: Where am I going? (feed up); How am I going? (feed back); and Where to next? (feed forward). Teacher educators work through each of these questions as they interpret the evidence generated from cross-institutional moderation.

TABLE 8.1 The relationship between professional standards and the GTPA at input and output in the feedback loop

	System input	*System output*
Role of professional standards	• Inform curriculum and program design • Embedded in GTPA design	• Judgment of GTPA • Analysis and reporting of moderation activity
GTPA activity: Data as feedback and feedforward	• Embedded across the program to sequentially develop relevant skills • Aligned with *Planning*, *Teaching*, *Assessing*, *Reflecting*, and *Appraising*	• Calibration, scoring, and moderation • Feedback o for credentialing purposes o for evidence of impact • Feedforward o to inform curriculum review and program renewal

Answering the question 'Where am I going?' requires a base from which to measure growth or change. In the case of TPAs, typically the starting measure of impact is the program design as this is informed by the system inputs which are the Australian Professional Standards for Teachers (APST) at the graduate level (Australian Institute for Teaching and School Leadership; AITSL, 2011) and the National Program Standards (AITSL, 2015). The standards, understood within their sociocultural context, provide a basis for program planning and the sequential development of knowledge, skills, and dispositions expected of graduating teachers. The generated moderation report provides an indication of 'How am I going?', connecting to the elements of the assessment: *Planning*, *Teaching*, *Assessing*, *Reflecting*, and *Appraising* (see Chapter 4). The report allows for structured feedback conversations informed by evidence. The value of this process is reliant on the development of trust and quality relationships between the research institute[3] (as the provider of the report) and teacher educators across participating universities in the GTPA Collective. Their interrogation of the reported evidence considered in teaching teams leads to program redesign and renewal, that is, 'Where to next?'.

Evaluative expertise and data literacy: Scoring teaching performance assessments

Evaluative expertise describes the ability to discern quality in a field of interest or practice. Consider, for example, judges that come together to identify the best wine or painting for an award who bring deep knowledge of the field and the features of quality relevant to the appraisal. More than that, experienced judges know how to apply the criteria and combine them to arrive at an overall judgment. This evaluative ability develops over time through practical experiences and engagement with colleagues and relevant artifacts (Wyatt-Smith & Adie, 2021b). In the literature, evaluative expertise is often used alongside performances recognized to be complex. The criteria for judging complex performances have been described as 'fuzzy', that is, where descriptive terms used to capture quality may lack precision and remain open to interpretation (Sadler, 1985). Sadler further portrayed such criteria as "deeply rooted in experience" (p. 293), explaining that judgment first occurs through an impression of quality. Wyatt-Smith and Adie (2021b) identified practices that promoted the development of evaluative expertise, for example, discussing criteria with colleagues, sharing and modeling how to use criteria with students, developing artifacts that explain a judgment process, and involvement in moderation.

In our experience of the GTPA, through induction and calibration training (see Chapter 7), scorers learn to apply common criteria to different performances and contexts, achieving high reliability (see Chapter 6 and below). Referring to the United States, Cohen et al. (2018) concluded that while systemic processes push for standardization of performance through reforms such as the edTPA, site-level responses must remain "as nuanced and complex as the system itself" (p. 22).

For this to occur, teacher educators require "the ability to ask and answer questions about collecting, analyzing, and making sense of data" (Hamilton et al., 2009, p. 47), otherwise referred to as 'data literacy'.

Mandinach and Gummer (2016) have provided a conceptual framework of data literacy for teachers in which they identified seven forms of knowledge required in data use. These include knowledge of content, curriculum, general pedagogy, subject-specific pedagogy, learners and their characteristics, education contexts, and educational ends, purposes, and values. Data use requires teachers to "identify problems and frame questions, use data, transform data into information, transform information into a decision, and evaluate outcomes" (p. 369).

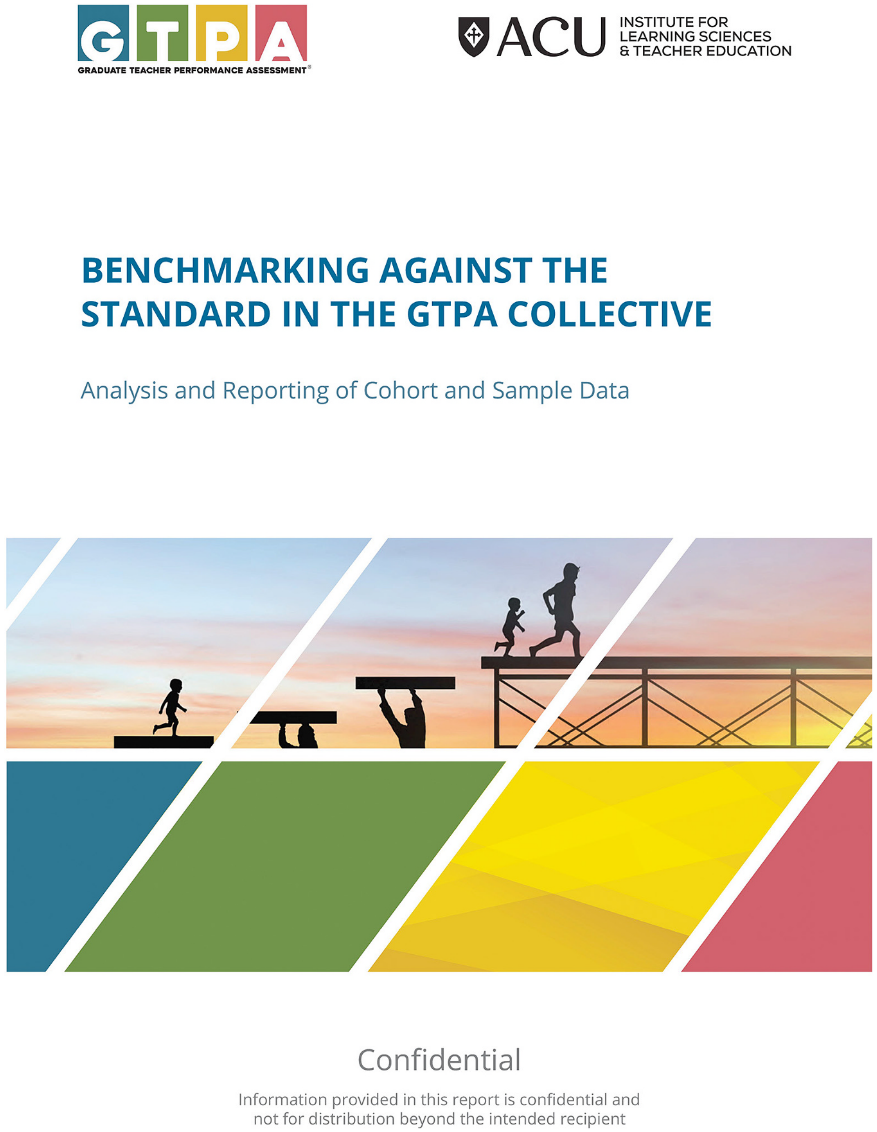
Teacher educators' knowledge, skills, and dispositions toward the collection and use of data have been shown to influence its effectiveness (Datnow & Hubbard, 2016). Teacher educators require the knowledge, skills, and confidence to use data, as well as a belief that data use to inform planning and teaching is part of their role. For effective data use in program renewal, teacher educators' trust in the validity of the source of evidence is essential as well as the willingness to accept and act on evidence (Cowie & Cooper, 2017).

In the GTPA project, the process of engaging teacher educators in data literacy and the use of large-scale data for evidence was made possible through a multidisciplinary research team with expertise, not only in teacher education, but also in high-end data analytics and digital infrastructure design. The building of data literacy was advanced during in-person and online meetings to canvass how the information from data analysis could be applied. Also canvassed were the types of data visualization most useful for showing GTPA scoring outcomes and related analyses. It was crucial that the teacher educators had input into the form in which the evidence from data analysis was returned to them. This was a critical element in enacting a collaborative partnership through teacher educators and researchers transforming "the numbers and statistics into instructional strategies that meet the needs of specific students" (Mandinach, 2012, p. 76).

The customized GTPA Report

The analyzed data from GTPA CIM-Online™ scoring is used in an automated report generated annually to provide confidential evidence to universities. The function of the customized report, *Benchmarking against the standard in the GTPA Collective*, is part of the evidence-informed feedback loop to universities to inform institution-level reporting and decision-making (Figure 8.2).

The report consists of two parts. Part One provides evidence showing how the university has applied the standard within and across programs. This gives insight into the comparability of the standard as applied to the corpus of samples. Part Two shows cohort data submitted by the university including patterns of performance at the criterion level. The submitted data includes judgment at the criterion level for all performances, and contextual information about the ITE program (e.g., mode of delivery and campus; student placement characteristics).



GTPA
GRADUATE TEACHER PERFORMANCE ASSESSMENT®

ACU INSTITUTE FOR LEARNING SCIENCES & TEACHER EDUCATION

BENCHMARKING AGAINST THE STANDARD IN THE GTPA COLLECTIVE

Analysis and Reporting of Cohort and Sample Data

Confidential

Information provided in this report is confidential and not for distribution beyond the intended recipient

FIGURE 8.2 Cover page of the customized report returned to each university in the GTPA Collective with analysis of sample and cohort data

The report is a significant source of information for program renewal within a university. Teacher educators need to learn how to read this information consisting of tabular and graphic visualizations and use it to inform decisions about improvements to their programs. They also need to learn how to combine the cross-institutional, moderated, criterion-level data in Part One of the report and the institutional program cohort or census data reported in Part Two. Information from the GTPA Report sits alongside the teacher educators' own analysis of their site-specific GTPA samples and other relevant program and cohort data. The two parts of the report are designed to be complementary and deliberately linked in interpretations of performance. Each of these parts is discussed in turn.

GTPA Report: Part One – Application of the standard

Part One presents information intended primarily to report on a summative judgment of Meets/Does not Meet the established standard. It includes:

1. Graphs of performance locations relative to the established standard to achieve an overall pass. These are derived from the analysis of CIM-Online™ scores using the multi-facet Rasch model method (MFRM; Figures 8.3 and 8.4).
2. Patterns of sample performance relative to the stated criteria (Figure 8.5).
3. Relative judgment severity of raters (Figure 8.6).

There are different ways of visualizing performance against a standard and relative placement. Figure 8.3 is one of the approaches that we have used in communicating the results to members of the Collective. Figure 8.4 presents an alternate visualization showing samples grouped together or co-located in bands. Both visualizations show relative rank order (horizontal axis) and the relative measure of performance for each GTPA sample on the logit scale (vertical axis).

Figure 8.3 shows each scored sample as a dot on a slope representing the lowest scored sample in the bottom left-hand corner to the highest scored sample in the top right-hand corner. The vertical dashed line represents the minimum acceptable level for meeting the standard. The position of a sample on the slope shows the ranked performance of the sample relative to the standard as determined by the

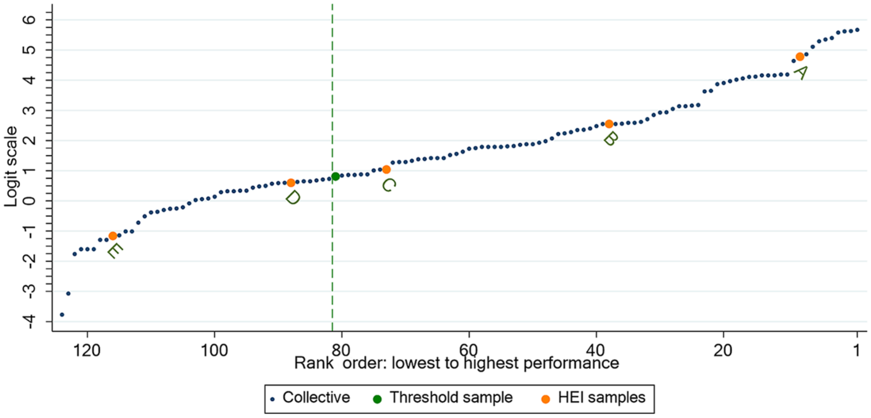


FIGURE 8.3 Relative sample performances in the total pool, with position highlighted for university samples (hypothetical data)

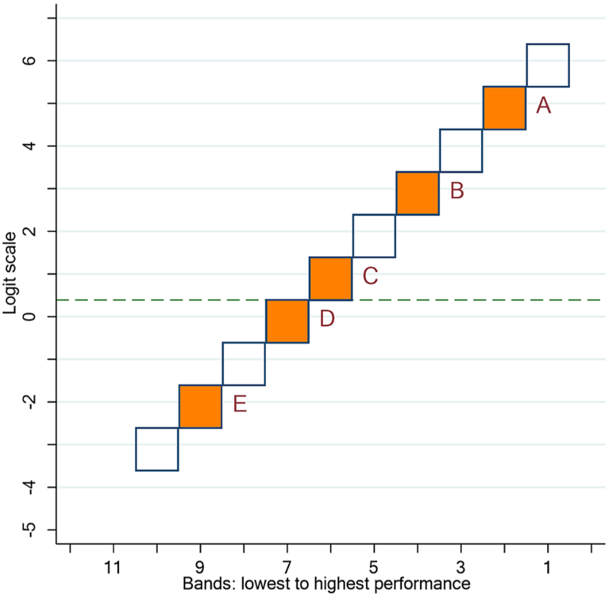


FIGURE 8.4 Relative sample performances in the total pool grouped in bands (hypothetical data)

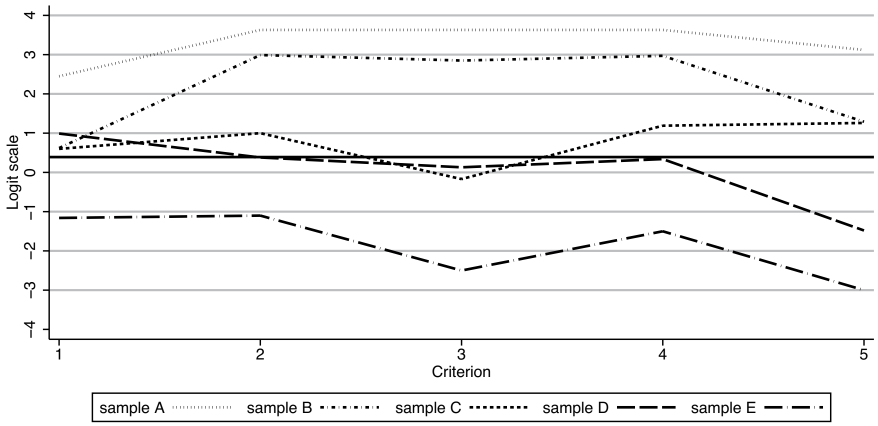


FIGURE 8.5 Ordered patterns of sample performance relative to the stated criteria, corresponding to the samples highlighted in Figures 8.3 and 8.4

Collective through the GTPA CIM-Online™ activity. A sample meets the standard if it is placed to the right of the vertical line on the slope. Conversely, a sample does not meet the standard if it is placed to the left of the vertical line. For example, in Figure 8.3, the Collective has determined that samples D and E do not meet the standard, and that samples A, B, and C do meet the standard.

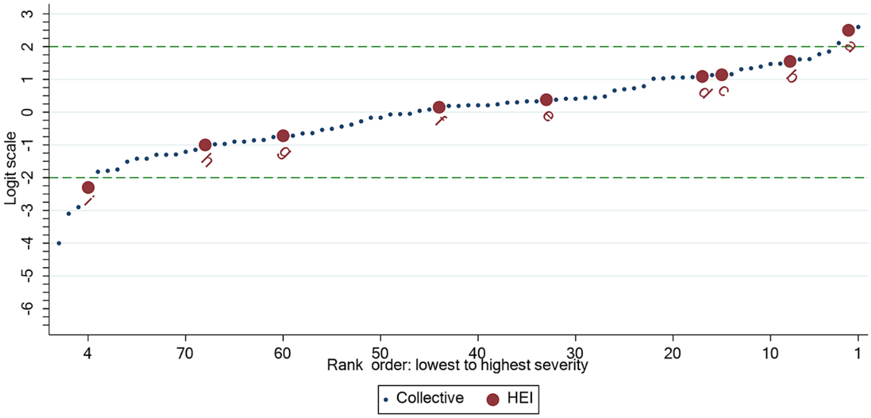


FIGURE 8.6 Measure of relative rater severity in scoring a GTPA for the Collective panel of judges, with university raters highlighted

Figure 8.4 shows the samples grouped together in bands (represented by a square) of similarly ranked samples. There are ten bands. Samples provided by the university receiving the report are located in the bands with black shading. The lowest-ranked samples appear in the bottom left-hand corner. The highest-ranked samples are located in the top right-hand corner. The height of each band represents two standard errors of the estimated sample location. The dashed horizontal line represents the minimum acceptable level on the scale. Bands above the dashed line represent groups of samples that meet the standard. Bands below the dashed line represent groups of samples that do not meet the standard. For example, in Figure 8.4 samples D and E do not meet the standard, while samples A, B, and C meet the standard.

The information in Figures 8.3 and 8.4 can be used by teacher educators within the university to compare their judgment of each submitted sample with the judgment of the Collective against the established standard. The university judgment of the sample will either be confirmed by the Collective through agreement, or an inconsistency in judgment will indicate to the university that their application of the standard needs to be reconsidered. This may be through further training of teacher educators on judging samples against the standard. If the Collective is consistently judging a university's samples to be below the standard, then this may indicate that a review of the program is required to improve the quality of preservice teachers' performance on the core skills measured through the GTPA.

To accompany the relative placement graphs (Figures 8.3 and 8.4), a line graph in the reports (Figure 8.5) is included to show how raters scored each sample on the five GTPA criteria (see Chapter 7 for a more detailed discussion). For a single sample, a line connects the estimated location on the common logit scale for the five criteria. In Figure 8.5, the criteria are represented on the horizontal axis (criterion 1 – *Planning*; criterion 2 – *Teaching*; criterion 3 – *Assessing*; criterion 4 – *Reflecting*; criterion 5 – *Appraising*). The minimum acceptable level for meeting the standard is shown by the horizontal solid line. For example, sample C is located above the solid line for criteria 1, 2, 4, and 5 and below the line for criterion 3. An area for immediate attention would be criterion 3 (assessing).

Figure 8.5 indicates patterns of performance in each sample across the criteria. It reveals criteria that are typically performed at relatively higher and lower levels within a program, revealing focus areas for review and improvement. For example, there appears to be an opportunity for focused attention on improving the skills of appraising in the program. The cohort data in Part Two of the report would be useful in this process.

It is well recognized that some raters ascribe to themselves the title of 'hard markers' as though this were a badge of honor. To complement the focus on judgment reliability, we focus on rater severity recognizing that overly severe or overly lenient raters could present risks to reliability. For this reason, the report provides information on the relative judgment severity of raters within the university. Each rater who participates in CIM is required to complete calibration training prior to scoring up to 15 samples (see Chapter 7). The MFRM analysis, referred to above, produces an estimate of judgment severity associated with each rater. This measure of judgment severity on the logit scale is plotted against the rank order of raters by judgment severity, from lowest to highest, as shown in Figure 8.6.

The horizontal dashed lines on the graph represent the lower and upper 95% confidence limits for judgment severity, respectively. Raters with severity measures that fall outside these limits are considered unlikely to have severity measures that are consistent with the distribution of the severity for other raters in the pool. For the example illustrated in Figure 8.6 nine raters from one university participated in the scoring activity from a pool of more than 80 raters. The nine raters are denoted by the letters a–i and are represented on the graph by the nine large solid dots. Rater 'a' has a severity measure that falls just above the upper limit of severity and hence, this rater was considered significantly more severe in their scoring of samples than the average rater. That is, rater 'a' was more likely to score their allocated samples as below the standard when other raters scored the same samples as meeting the standard. Conversely, rater 'i' has a severity measure that falls below the lower limit of severity and hence, this rater was considered significantly more lenient in their scoring of samples than the average rater. That is, rater 'i' was more likely to score their allocated samples as meeting the standard when other raters scored the same samples as below the standard. This information can be used by the university to identify that additional training is required to build rater dependability in their team.

GTPA Report: Part Two – Analysis of program cohort data submitted by the university at the criterion level

Part Two reports outcomes of the analysis of program cohort data including patterns of performance at the criterion level. There is a formal requirement that this data has been internally moderated before submission. The data, submitted by each university, include performance scores against each of the five criteria (see Chapter 4).

Additional data are collected including contextual information gathered to give a situated perspective on individual and cohort performances. Information about the ITE program type is collected as well as school placement characteristics. Program characteristics of interest are: undergraduate or postgraduate degree, teaching focus of the program (early childhood, primary, secondary), learning area (discipline/subject), mode of delivery (on-campus, online, blended), and campus or location of delivery. School placement information includes phase of schooling, that is, school year level and school location (metropolitan, regional, remote). The analyses of these data for the whole program cohort are useful for examining program effectiveness and impact.

Data visualizations and descriptive analyses of performance on the stated criteria are used to represent program characteristics. Figure 8.7 is an example of one type of data visualization provided in the second part of the report. It uses percentages to convey comparative performance on each criterion (C1–C5) at the cohort level across a suite of programs. The information enables the identification of those criteria that were more or less difficult for the cohort to achieve.

The analysis of performance at this level is designed to be used for curriculum planning and program renewal. For example, Criterion 2 is an identified area of underperformance in the Master of Teaching (Secondary); Criterion 4 is an identified area of underperformance in the Bachelor of Education (Primary);

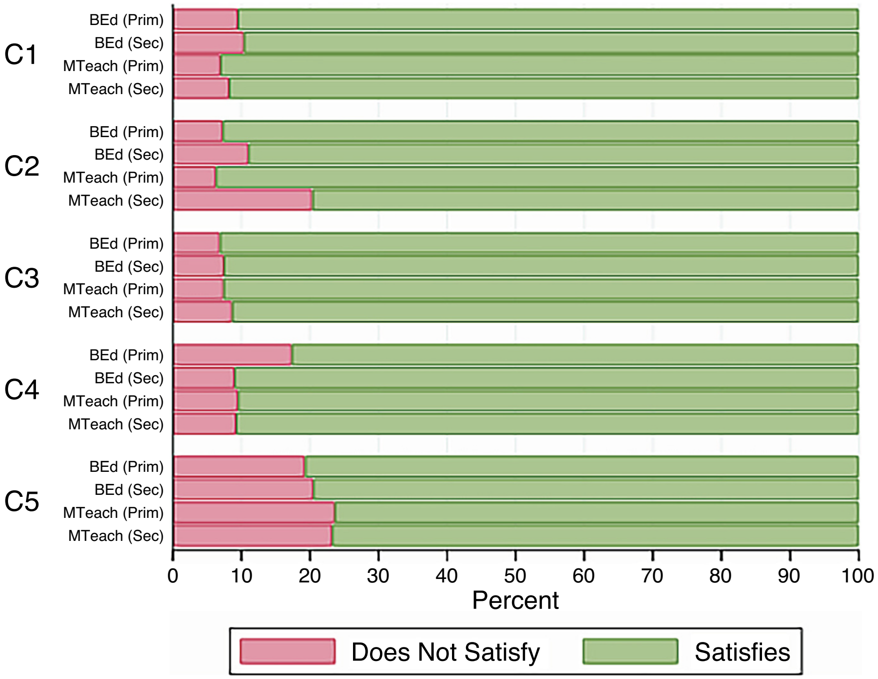


FIGURE 8.7 Bar graph showing percentages of GTPA performances that Do Satisfy and Do Not Satisfy the criteria requirements as applied to each program

and Criterion 5 is an identified area of underperformance across all programs. This evidence highlights areas for teaching focus and improvement within identified programs.

The value of feedback loops that are informed by data from multiple institutions

Our journey to investigate the potential of TPAs has led us to understand the function of standards and evidence as enabling when they focus on the use of data to build capacity and make informed decisions. This has opened new, evidence-informed ways of seeing and talking about teacher education. The GTPA Reports bring into scope two main types of data: externally moderated performance data assessed against an established standard, and internally moderated cohort performance data at the criterion level. Taken together these provide ways to see program quality in the achievement of the standard. We use the phrase 'endorsement of judgments' to refer to the agreement between the score initially awarded by the university and the finalized score from CIM. Where there is a discrepancy between the two, this is highlighted in the report and discussed in a dedicated meeting addressing the dependability of judgment and approaches to internal moderation. Further, comparative performance of a program over time, assessed against the standard, is discussed as providing a longitudinal measure of program quality.

Through using the CIM data, the research team and teacher educators can identify and discuss quality in five areas of teaching practice within and across programs. A methodology for spotlighting areas for improvement has been developed through the integration of CIM analyzed data and the cohort level criterion scores. For example, as identified previously, the university receiving the graphs in Figure 8.7 based on cohort level criterion scores, could identify appraising the effectiveness of practice (criterion 5) as an area for improvement across all programs. Those teaching into the Master of Teaching (secondary) program could also investigate how subject-specific pedagogies (criterion 2) are being incorporated, and those teaching into the Bachelor of Education (Primary) program, could also investigate how reflection (criterion 4) is taught and developed in that program. However, these data need to be interpreted with reference to the program-specific results of CIM and the endorsement of judgments made on selected samples within programs. Developing data literacy to read and effectively use the GTPA Reports is thus an essential skill for teacher educators.

Different universities can receive different levels of endorsement dependent on a variety of factors such as the familiarity of the teacher educators and their prior induction into the GTPA. To illustrate, using hypothetical data, Figure 8.8 shows the predicted probabilities of endorsement from a logistic regression analysis, including 95% confidence intervals (CI) by university and taking sample quality into account. The horizontal line indicates the overall endorsement with the 95% CI represented as the horizontal dashed lines. Statistically significant differences were found particularly between university 2 and universities 6–12; universities 6–12

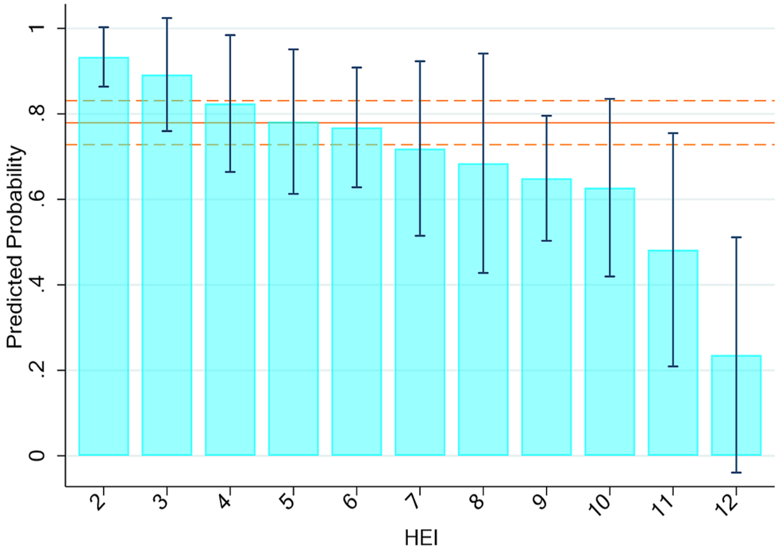


FIGURE 8.8 Predicted probabilities of endorsement of judgments across 12 universities: Results from a logistic regression analysis

were less likely to have judgments endorsed compared to university 2. This variation was proven to be statistically significant, even when sample size and sample quality were taken into consideration. A variation across universities in endorsement implies that the common standard is not being consistently applied across all programs. This identifies an area for inquiry within a university on how to 'fill this gap' between the actual and intended or desired level of endorsement.

The aim of calibration is for all judges to have a shared knowledge of the standard and be able to recognize and apply this in actual GTPA performances. The importance of a shared understanding of the standard, prior to participating in CIM, is evident in Figure 8.9. As participation in calibration has increased over 2018–2020, the overall endorsement of GTPA samples selected for CIM has increased to 80% in the 2020 GTPA implementation. Calibration not only prepares judges for CIM, but also enables the monitoring of the judges' competencies in the application of the standard.

Through the collated data, we can see changes occurring in programs in the short term. Table 8.2 shows another way to see endorsement over time. In the table, a competency rank has been given to a subset of universities in the GTPA Collective. "Gold' represents an endorsement of 90% or greater; 'silver', endorsement between 80–90%; and 'bronze', endorsement below 80%. In 2018, three out of five universities were ranked 'bronze', while in 2020, four out of five were ranked 'gold'. All but one of these universities show increased competency in applying the standard across this period. We can see from this visualization, improvements in understanding and recognizing the standard over time. University 9 which remained at a 'bronze' level

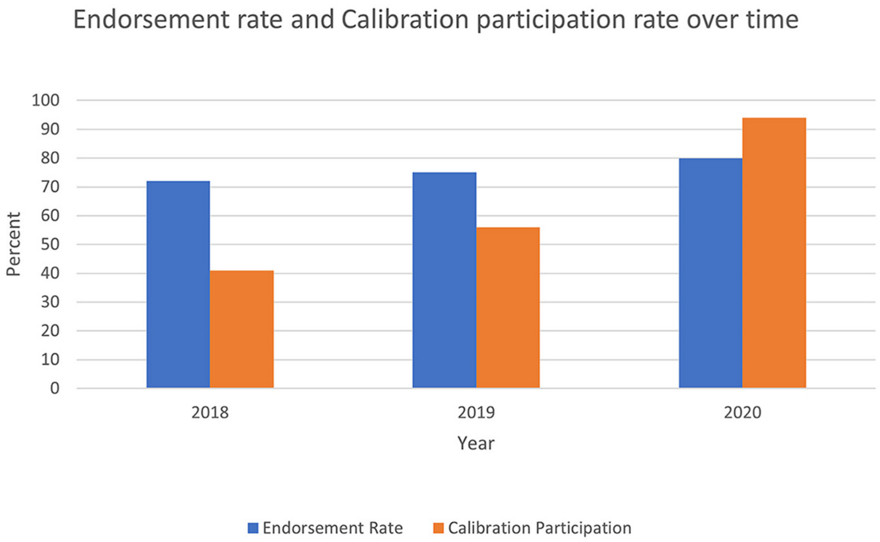


FIGURE 8.9 Endorsement rate and calibration participation over time (2018, 2019, 2020)

TABLE 8.2 University ranking of applying the standard as a unit based on endorsement: gold ≥ 90%; silver = 80–90%; bronze ≤ 80%

HEI	*2018*	*2019*	*2020*
1	Bronze	Gold	Gold
2	Bronze	Gold	Gold
3	Silver	Silver	Gold
5	Silver	Bronze	Gold
9	Bronze	Bronze	Bronze

of endorsement could enquire into the mitigating factors. For example: Has there been a changeover of staff? Do internal calibration and moderation processes need to be strengthened?

We have learned that alongside a validated TPA is a need to develop expertise that is research- and evidence-informed. Developing evaluative expertise in the GTPA Collective has occurred through immersion in a national community willing to learn and inquire into practice using evidence. The establishment of a feedback loop (Figure 8.1) supports the professionalism of teacher educators as a self-regulating community, foregrounding intelligent accountability in the use of data (see Chapter 3). The feedback loop provides information about the criteria, the standard, calibration participation, and the endorsement of judgments at the program level to see the impact. Brought into focus is a new relationship between formative and summative purposes of assessment in teacher education.

Conclusion

The data in the GTPA Reports show, for the first time, evidence of the preparedness of preservice teachers graduating from Australian ITE programs assessed against an established standard. The chapter presents evidence of teacher educators' increasing engagement in calibration and cross-institutional moderation, both foundational to building dependability of judgment. Teacher educators now have the evidence to make statements about the effectiveness of their programs and their impact on preservice teacher learning. Readers are referred to Chapter 11 to see how we are profiling progress in the GTPA with a focus on implementing an established standard. Also introduced are new investigations underway in workforce studies that follow graduates from entry into teacher education through to graduation and into the teaching workforce.

Notes

1 Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

2 Acknowledgment: The online model of cross-institutional moderation (CIM-Online™) was conceptualized and developed in the Institute for Learning Sciences and Teacher Education, Australian Catholic University. For a discussion of CIM-Online™, readers are advised to also see Wyatt-Smith and Adie (2021a).

3 The GTPA is hosted by the Institute for Learning Sciences and Teaching Education (ILSTE), Australian Catholic University (ACU).

References

Australian Institute for Teaching and School Leadership (AITSL). (2011). *Australian professional standards for teachers*. https://www.aitsl.edu.au/teach/standards

Australian Institute for Teaching and School Leadership (AITSL). (2015). *Accreditation of initial teacher education programs in Australia: Standards and procedures*. https://www.aitsl.edu.au/docs/default-source/default-document-library/accreditation-of-initial-teacher-education-programs-in-australia_jan_2019.pdf?sfvrsn=4639f33c_2

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Cochran-Smith, M. (2004). Defining the outcomes of teacher education: What's social justice got to do with it? *Asia-Pacific Journal of Teacher Education*, *32*(3), 193–212.

Cohen, J., Hutt, E., Berlin, R. L., Mathews, H. M., McGraw, J. P., & Gottlieb, J. (2018). Sense making and professional identity in the implementation of edTPA. *Journal of Teacher Education*, *71*(1), 9–23. https://doi.org/10.1177/0022487118783183

Connell, R. (2009). Good teachers on dangerous ground: Towards a new view of teacher quality and professionalism. *Critical Studies in Education*, *50*(3), 213–229. https://doi.org/10.1080/17508480902998421

Cowie, B., & Cooper, B. (2017). Exploring the challenge of developing student teacher data literacy. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 147–163. https://doi.org/10.1080/0969594X.2016.1225668

Darling-Hammond, L. (2008). Reshaping teaching policy, preparation and practice: Influences on the National Board for Teaching Professional Standards. In L. Ingvarson & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 25–53). Emerald.

Darling-Hammond, L. (2012). The right start: Creating a strong foundation for the teaching career. *Phi Delta Kappan*, *94*(3), 8–13. https://doi.org/10.1177/003172171209400303

Datnow, A., & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change*, *17*(1), 7–28. https://doi.org/10.1007/s10833-015-9264-2

Delandshere, G., & Arens, S. A. (2001). Representations of teaching and standards-based reform: Are we closing the debate about teacher education. *Teaching and Teacher Education*, *17*(5), 547–566. https://doi.org/10.1016/S0742-051X(01)00013-0

Forde, C., McMahon, M., Hamilton, G., & Murray, R. (2016). Rethinking professional standards to promote professional learning. *Professional Development in Education*, *42*(1), 19–35. https://doi.org/10.1080/19415257.2014.999288

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making* (NCEE 2009-4067). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/wwc/Docs/PracticeGuide/dddm_pg_092909.pdf.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, *47*(2), 71–85. https://doi.org/10.1080/00461520.2012.667064

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, *60*(November 2018), 366–376. https://doi.org/10.1016/j.tate.2016.07.011

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13. https://doi.org/10.1002/bs.3830280103

Sadler, D. R. (1985). The origins and functions of evaluative criteria. *Educational Theory*, *35*(3), 285–297.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*(2), 119–144. https://doi.org/10.1007/BF00117714

Stanford Center for Assessment, Learning & Equity. (n.d.). About edTPA: Overview. http://edtpa.aacte.org/about-edtpa#Overview

Talbot, D. (2016). Evidence for no-one: Standards, accreditation, and transformed teaching work. *Teaching and Teacher Education*, *58* (August), 80–89. https://doi.org/10.1016/j.tate.2016.05.006

Wyatt-Smith, C., & Adie, L. (2021a). Introducing a new model for online cross-institutional moderation. In C. Wyatt-Smith, L. Adie, & J. Nuttall (Eds.), *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration* (pp. 39–58). Springer.

Wyatt-Smith, C., & Adie, L. (2021b). The development of students' evaluative expertise: Enabling conditions for integrating criteria into pedagogic practice. *Journal of Curriculum Studies*, *53*(4), 399–419. https://doi.org/10.1080/00220272.2019.1624831

Wyatt-Smith, C. M., & Bridges, S. (2008). Meeting in the middle–assessment, pedagogy, learning and students at educational disadvantage. Final Evaluation Report for the Department of Education, Science and Training (DEST) on Literacy and Numeracy in the Middle years of Schooling. DEST.

9

CULTURE CHANGE IN INITIAL TEACHER EDUCATION

How shall we know quality and impact?

Introduction

Change is a constant in contemporary times. We see and live change constantly; it is pervasive, shaping in profound ways how we come to know and understand ourselves and the world around us. This holds true for education, schooling, and learning more generally. The roles of teachers and our understanding of teacher professionalism are undoubtedly in a state of flux, reflecting societal and economic changes and the juggernaut of digital technologies. In this context, it is hardly surprising that policy has sustained and even intensified its laser-like focus on quality teaching and the relationship between teaching and learning outcomes. Building on the observation that the terms *quality* and *impact* are far from stable and remain ill-defined in policy, practice, and research, our interest is in how we talk about overdue changes to teacher professionalism and what this means for the work of teacher educators.

As discussed in earlier chapters, the Teacher Education Ministerial Advisory Group (TEMAG) review of initial teacher education (ITE) in Australia identified the need for "structural and cultural change" (Craven et al., 2014, p. xii). This included strengthening partnerships between universities and schools, and rigorous authentic assessment to produce "robust assurance" of preservice teachers' readiness for classroom practice (p. xiii; see Chapter 2 for a discussion of readiness), along with other changes in preparation. This chapter explores the experiences of teacher educators, recent graduates, preservice teachers, policy personnel, mentor teachers, and other school leaders who have been directly involved in implementing the Graduate Teacher Performance Assessment (GTPA®).[1] Particular attention is paid to the phenomenon of how a purpose-designed community of teacher educators

DOI: 10.4324/9780429318504-11

embraced change, not as an end in itself, but rather with the intent of improving graduate preparedness and in turn, the learning outcomes of Australia's young people.

A corpus of recorded talk and interactions is drawn on to offer illustrative examples of working with diverse universities and other related organizations (e.g., teacher regulatory authorities responsible for program accreditation and members of peak professional associations representing teachers and school leaders). Throughout, there emerged a dominant discourse of teaching as a subject of inquiry putting student learning at the center. Also evident was the growing attention given to the use of data and evidence to appraise and reflect on the impact of teaching practice on learning. For the research team, we adopted the position that it was necessary to be attuned to developments and experiences as they happened locally or in situ, and while engaging across contexts and at scale. Our experiences highlighted how new knowledge and expertise carry forward through a culture of collaboration. So, how was this attempted?

Collaborative professionalism and professional identity

As described in the field of business, collaborative professionalism encapsulates the interdependencies within professional work that involves shared exchanges of knowledge and draws on the specialized expertise within a team (Adler et al., 2008; Racko et al., 2019). Referring to schooling, Hargreaves and O'Connor (2017) use the term collaborative professionalism to describe cases of teachers working together with collective responsibility for improvement through the interrogation of evidence. The GTPA project has been informed by a conceptualization of collaborative professionalism, as well as by Sachs' (2001) description of two discourses of professionalism: *managerial* which is embedded in quality and accountability agendas, and *democratic* in which the profession takes control of itself through inquiry into its own practice. In the GTPA Collective, our common aim is to undertake investigations into ITE practice for improvement and reporting purposes. The introduction of TPAs in Australia was undoubtedly a policy-driven reform with the new assessment intended to support accountability agendas. While the GTPA research team recognized this from the beginning, we deliberately chose to position the work within a democratic discourse. This intentional reframing of teacher educators' work involved a collective inquiry into practice, working with teachers in schools and policy personnel.

A main challenge was that the teacher education community had little, if any, sustained experience of working together to share multiple perspectives about program quality and professional competence. To build a safe space for sharing learning and developing trust, the research team decided to make an explicit provision for the GTPA Collective to meet regularly. This involved in-person and online meetings and events, including whole of group online monthly meetings known as Touchpoint Sessions. Additionally, there were formally arranged individual and group meetings and others that were incidental. These occurred as the need emerged, for example,

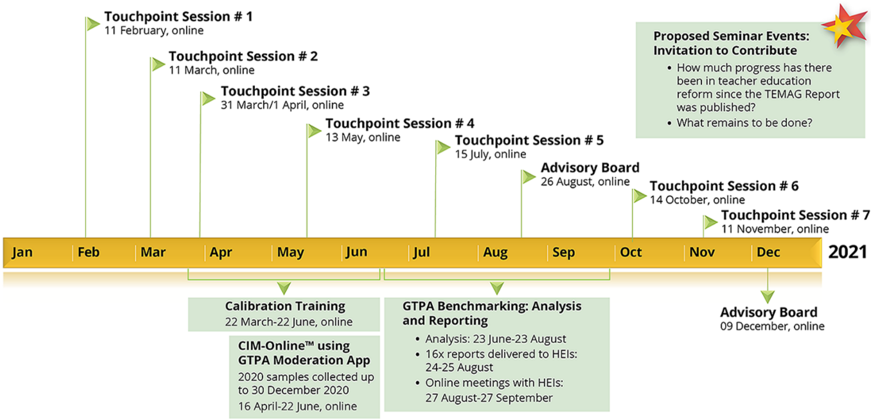


FIGURE 9.1 2021 schedule of GTPA collective meetings and events

as part of onboarding new university members into the Collective. See Figure 9.1 for an example of an annual schedule of GTPA meetings and events. Our intent was to open the space for voices from diverse settings and institutional contexts to share theoretical and practice-based knowledges and expertise, and interrogate what makes an authentic complex performance assessment. An early question, for example, was how such an assessment was similar to, and different from, other assessments in the ITE program, and what would count as 'ready to teach' evidence. The collaborations recognized the centrality of disciplinary and evaluative expertise, state policy and regulatory requirements, and the full range of university preparation programs from early years to senior schooling.

The expanded relationships have provided access to different perspectives to see and advance inquiry through a range of 'eyes' (ways of seeing). Further, the new policy space of performance assessments has provided an opportunity to try-on new types of agency in ITE, with the potential to shape new identities for teacher educators, preservice teachers, policy officers, teachers, and researchers. Even beyond this, and perhaps more importantly, it provided a stimulus to advance the focus of inquiry from *What am I going to teach?* to *How will I know what, how, and how well students are actually learning?* As a group, this refocusing challenged our thinking about how to use data as evidence to inform reflection on the quality and impact of teacher education programs.

As suggested earlier, professional identities were being formed and reformed across contexts through collaboration. The reformation of professional identity broadened from an institutional identity to include alignment with the larger Collective, drawn from across the country, offering diverse experiences and accounts of ITE. In this space, the teacher educators took up the role as change agents to lead the introduction of the GTPA within their own institution as well as contributing to collective knowledge through their developing expertise of implementing, scoring, and moderating results from the GTPA (see Chapter 7).

Prioritizing teacher educator agency alongside collaborative action was essential to counter a widely perceived, even dominant, view of a TPA as a technicist tool intended to standardize ITE (see Chapter 3). Working from this position, some teacher educators experienced dissonance when the requirements of the new assessment appeared to be different from more traditional and already accepted approaches with which they were most familiar. These had been previously forged and normalized as routine in their jurisdictional and geographic contexts and in disciplinary/subject content and program teams. Opportunities for social and authentic engagement and discussion of shared opportunities and challenges were critical in this experience of change. As shown in the discussion of talk segments that follow, new identities were being forged as teacher educators shared their experiences of implementing and scoring the GTPA individually and with others. In this process, they were also introduced to new and different methodologies that promoted inquiry into issues of standards, evidence of quality, judgment, and assessing competence more generally. This social process was integral to how teacher educators were individually and collectively active in interpreting their experiences (Colmer, 2017) and moving to adopt shared responsibility for teacher education reform.

Analyzing teacher educator talk and interactions

The data discussed in this section draws on 52 hours of recorded talk. Participants included teacher educators from the GTPA Collective, recent graduates, preservice teachers, policy personnel, mentor teachers, and other school leaders. The talk and interactions (N = 41) occurred over five years (2017–2021). Talk was recorded and transcribed in full. University ethics approval was obtained for the collection and use of the transcripts for analysis and academic publications. Analysis was undertaken using NVivo 12 qualitative data analysis software (QSR International, 2018) to identify dominant and recurring themes across the talk and interactions. Table 9.1 profiles the corpus of talk data analyzed for this chapter.

Silverman's (2006) and Freebody's (2003) ethnomethodological approaches to discourse analysis were drawn on. The transcripts were read and reread by multiple members of the research team. We examined the points of consonance and dissonance in the talk to hear accounts of changes in identities and work practices triggered by the mandated introduction of TPAs. Throughout, the focus was sustained on what the talk and interactions revealed about the identity of the speaker: the

TABLE 9.1 Corpus of talk data

Event	*Date/period*	*Number of participants*	*Hours of talk*
Touchpoint Sessions	2017–2021	428	40.0
Interviews	2018, 2020–2021	14	9.5
Conference presentations	2019	19	2.5
Total	**2017–2021**	**461**	**52.0**

ways of thinking, being, acting, and feeling made available in the transcripts. Guided by the understanding that over time, discourses serve to naturalize values, beliefs, and social practices, our interest was in the ways of talking that participants relied on to talk about their experiences of change associated with the GTPA, and what this meant for previously taken-for-granted ways to prepare teachers. Of interest was what the talk revealed about shifts in understandings of the work of teacher educators and the openings created by the GTPA to consider the introduction of new practices and relationships. The analysis of talk makes it possible to capture understandings and beliefs of individuals and groups as they come together into new social groupings.

A TPA is not a thing; it is a concept. It takes time and appropriate conditions for agreement to emerge in an ITE community about what it stands for. Working from this position, how the assessment is designed and relates to standards is but one focus of inquiry; the other equally important focus is how those involved understand the nature and function of the assessment and how they experience change. This conceptualization supports inquiry going beyond policy and practice changes to changing identities. Transcripts of talk were examined to identify distinctive responses to change, producing five categories of response:

1. Conserving practice.
2. Aligning practice.
3. Enhancing practice.
4. Bringing judgment and moderation into sharp focus.
5. Realizing the potential of TPAs.

The categories are not offered as sequential or common across all teacher educators. Individuals could, for example, move to and fro across the five, in response to particular influences on their practice and thinking at local and system levels. As a provisional framing, they open out participants' ways of thinking about themselves – identities – and their relationships with each other in a time of intense change. They serve to capture how transformation was experienced by individuals and groups. They are not offered as distinct responses between which there are hard boundaries. Rather, as you listen to the voices, you will necessarily hear some overlap. The taking of shared responsibility and agency for change is striking, with teacher educators identifying areas for concerted actions and change, and preservice teachers experiencing how these shaped their emerging professional identities and practice.

The first response: Conserving practice

In the year-long trial of the GTPA, the dominant discourse taken up by teacher educators was of TPAs as instruments of standardization and compliance. From this vantage point, change was being imposed on them and was beyond their control. This reflected how the introduction of TPAs in Australia was part of a policy-initiated reform (see Chapter 1 for background). As can be heard in the teacher educator's talk below, the GTPA was perceived by many to be "just another assessment" and the main goal was to "tick the box". Teacher educators understood that they were to meet the official system requirement for demonstrating performance against professional and program standards (Australian Institute for Teaching and School Leadership [AITSL], 2011, 2015).

> The GTPA is seen as being a compliance issue rather than a program opportunity. So, we tick the box, we're done; we've got that, it meets the requirements of the standards. And this … gets reflected in statements that I hear around the traps like … it's just another assessment. Just like all the other assessments, it's not that big a deal.
>
> *(Teacher educator, Interview, 2020)*

This segment echoes the talk of many teacher educators unfamiliar with the concept of a performance assessment to establish profession readiness (see Chapter 2). In the trial, discussions concentrated on how to embed the assessment into the ITE program. This involved talk about where the assessment would fit into the program (timing), noting it was officially required to be a final year summative assessment for licensure. Decisions about effective processes for embedding the assessment remained with teacher educators: there was no official advice on how embedding was to occur over the course of the program. University assessment policies and individual preferences for assessment practices (formative and summative assessment) were perceived by many as needing to make space for the new assessment.

Initially, there was resistance to replacing existing traditional assessments, many of which had become normalized over time as to how teacher education is properly conducted. Embedding the assessment was talked about by some as "tacking on" the GTPA to already approved assessment tasks. For those teacher educators who used a portfolio method (e.g., a collection of evidence throughout the program), the accommodation was to expand the portfolio to include the GTPA as an additional item. This required preservice teachers to collect and collate different forms of evidence across their school-based practical program. In the words of one teacher educator, change was to be minimal: "In the current subject we're going to trial it in a 60% portfolio, and they also do a resource kit … and this will all be included in the GTPA" (Teacher educator, Touchpoint Session, 2017). In this talk segment, portfolios are collections of evidence with some commentary relating it back to relevant professional standards.

The GTPA was distinguishable by design from traditional understandings of portfolios as collections of evidence over time. It was conceptualized to elicit demonstration of *Planning*, *Teaching*, *Assessing*, *Reflecting*, and *Appraising* (see Chapter 4). It requires preservice teachers to present an evidence-informed demonstration of practice, including actual examples of classroom practice, decisions, and reasoning. It was designed to be more than an account of doing. It required teacher educators to shift their understanding of what was being assessed and how. It required that they understand competence assessment to establish profession readiness on program completion. However, most teacher educators who adopted a traditional mindset read the assessment as inviting the traditional "5000-word essay" response. They misread the intention to show competence in practice and decision-making by using evidence and linking theory and practice.

Where this misunderstanding occurred, teacher educators reported experiencing a sense of dissonance around the official expectation of their work. Some reported that they were required to provide formative feedback on all assessments to ensure preservice teachers "pass the test". The recurring question was how to provide feedback on a competence assessment intended to show that the preservice teacher was ready to teach. In the words of one teacher educator, the new assessment "sits interestingly philosophically with other aspects of the way that we do assessment" (Teacher educator, Touchpoint Session, 2017). Also at play were perceptions of practice, individual dispositions (Bair, 2017), and belief systems that typically colored pedagogies and local learning and assessment policies.

> It actually would break our teaching and learning policy here at the university to do what you're asking us to do. We could probably write our way around it, but my point is, there's a policy structure and that would not sit well within it.
>
> *(Teacher educator, Touchpoint Session, 2017)*

Teacher educators reported that their sense of discomfort became acute where it concerned the grading of preservice teacher performance as Meets or Does not Meet. Such responses about pass/fail were expected, given that the required standard of Meets had not been established previously (see Chapter 6). In the words of one teacher educator: "We're actually experiencing some difficulties around being able to meet internal grading requirements and the pass/fail of the GTPA" (Teacher educator, Touchpoint Session, 2018).

In the early years of GTPA implementation, and in particular, during the trial, preservice teachers were often reported as describing the assessment as 'quite overwhelming' and feeling 'panicked', 'scared', or 'anxious'. This too was not unexpected given that the competence assessment was a new requirement and in a form that was new to many. Further, earlier research had shown that the core tenets of assessment and the use of data as evidence to inform teaching and learning were not well established in teacher education programs (DeLuca & Johnson, 2017; Donaldson, 2010; Wyatt-Smith et al., 2017). The absence of such features in ITE programs is reflected in the description of the assessment by a recent teacher education graduate as "a big learning curve" (Interview, 2018). As can be heard in the segments below, there were also reports from preservice teachers that once the assessment had been completed, they reflected on the GTPA as being immensely 'valuable', 'helpful', and 'really worthwhile'.

> [I was] relieved that it was over, but at the same time, I found it very valuable because I got a glimpse of what teaching's like in comparison to just being at uni and learning theory.
>
> *(Teacher education graduate, Interview, 2018)*

> I felt overwhelmed at the beginning because it is very content-heavy, but once you broke it down into smaller parts, you were able to see the links … [to] the teacher standards [and] that's something I was really familiar with … it is a holistic approach of everything that you have learnt during your degree.
>
> *(Teacher education graduate, Interview, 2018)*

Preservice teachers were reported as describing the test as "a really difficult task, but it was the best one we've done" (Interview, 2018). One teacher educator described the GTPA as "really very relevant to them [preservice teachers]. They found that it was really good, and worthwhile. But it was hard, they had to work really hard at it" (Teacher educator, Touchpoint Session, 2017).

The first response revealed a level of uncertainty experienced by teacher educators and preservice teachers as they sought to implement what many regarded as an *alien assessment*. This reflects how the assessment called for a new metalanguage of teacher education that included talk about evidence and impact of practice. There was also a strengthened focus on assessment terminology such as validity, reliability, assessment fidelity, and integrity. Some experienced this shift as uncomfortable. As discussed in the following sections, teacher educators were also introduced to new scoring processes where their own judgment became a source of deep inquiry and reflection. Most teacher educators had not subjected their scoring processes to scrutiny, nor had they shared these with others in a large group setting of more than 80 teacher educators. In considering this first response, what is clear is the pragmatic approach taken by teacher educators whose priority was to graft the new assessment requirements onto the practices with which they were most familiar. This response was short-lived and changed as evaluative experience developed within the Collective.

The second response: Shifting mindsets to discern coherence and make connections

Following the one-year trial of the GTPA, there was a discernable shift away from a dominant discourse of pragmatism and compliance to a discourse of alignment. This shift was evident in the talk as participants described (1) aligning the practical and theoretical components of teacher education and bringing these together to see practice as a whole – this was related to developing a professional identity for preservice teachers – and (2) aligning the scope and sequence of teacher education to ensure opportunities for learning core planning, teaching, and assessing skills are provided.

Aligning the practical and theoretical components of teacher education

Teacher educators talked about the GTPA providing opportunities to rethink their program, planning, and teaching. They identified it to be a culminating demonstration of capability in which preservice teachers show *how* they bring together learning from both the academic program and the school-based practical program. In this way, the GTPA was identified by teacher educators to be located at the "nexus of theory and practice". The talk segments below suggest the GTPA functions as an integrating device:

> … a culminating assessment task that shows what [preservice teachers] have learned throughout the four years … what it means to be a teacher, how to work with students, how to differentiate instruction, how to assess students, how to build lessons that are practical lessons based on student data; all of these types of things … show this is what you need to do to be a quality teacher.
>
> *(Dean, Interview, 2020)*

> … it brings together theory and practice really well. Students have to think intentionally about what they're doing and why they're doing it and explain that explicitly. So, as a culminating task … in terms of professional readiness, I think it's very appropriate.
>
> *(Teacher educator, Conference Presentation, 2019)*

This perspective was shared by graduates and school leaders. A recent teacher education graduate stated: "it allowed me to put everything I've learnt into practice, and I was able to do that in the classroom" (Interview, 2018). A deputy principal reported that the GTPA "removes that box ticking compliance aspect that I have noticed previously, and it brings in that this is a whole story, a whole picture" of teaching and learning (Interview, 2018). He elaborates on this view in the segment below where he highlights the significant role of the GTPA as an authentic, culminating assessment of teaching practice that "brings together all of the aspects of teaching".

> I think the GTPA is really important in terms of bringing everything together. From my experience of preservice teachers, before the GTPA … they were just box ticking, standard by standard: "I've done this standard; tick. I've done this standard; tick" … and they wouldn't understand that it's a continua … almost like a circle that everything fits within. There's not really one bit of teaching that you can go, "that can just stay on that side" and "I don't really need to worry about that", "that's not important". So, it brings together all of the aspects of teaching. Teaching is not just one thing; it's a great many things.
>
> *(Deputy Principal, Interview, 2018)*

Developing a professional identity

Teacher educators described the GTPA as an opportunity to develop preservice teachers' professional identity and readiness for teaching. It was described by some as an "identity affirming experience that pushes preservice teachers to ask themselves the question: Am I ready?" (Teacher educator, Interview, 2020). One teacher educator reported that the assessment "helps preservice teachers to demonstrate professional practice in a metacognitive way" (Teacher educator, Interview, 2020).

In their reported talk, there are recurring instances where the notions of metacognition, identity, and professional responsibility are intertwined.

> If the professional experience assessment is where a supervising teacher can tell us they 'look' like a teacher, the GTPA enables us to say they 'sound' like a teacher in terms of what they're thinking, what they're saying about their teaching, and their decision making.
>
> *(Teacher educator, Interview, 2020)*

> … [professional experience] tells us whether you 'look' like a teacher, the GTPA tells us whether you sound like a teacher. And in doing so, the preservice teachers who are successful, they know that they 'sound' like a teacher from doing the [GTPA], we don't actually have to mark it for them to know that. So, in the process of … doing the GTPA, they start to recognize that they are in fact 'sounding' like the teacher. And they're showing their practices in a way that [says] I'm ready to do this and they're recognizing that readiness for themselves.
>
> *(Teacher educator, Interview, 2020)*

We hear the perspectives on 'becoming' a teacher: learning to look like a teacher and sound like a teacher. In the segments above, the speakers suggest the complementarity of school-based professional experience assessment and the GTPA assessment – the former scored by the supervising teacher and the latter by teacher educators. Further, the speakers make the link between 'sounding' like a teacher in terms of thinking, teaching, and decision-making. Recent teacher education graduates emphasized their increased employability, with many describing the assessment as providing them with the knowledge and skills to communicate their readiness for practice: "I've got the evidence and the proof… it's really helped me feel prepared where I wasn't feeling prepared before" (Interview, 2018). The key observation overheard in these segments is that in completing the GTPA, preservice teachers themselves come to recognize their emerging teacherly identity and stance in the classroom to the point of showing practice in ways that attest to their readiness and recognize this for themselves.

> This is a really high-level task, and even students [preservice teachers] said it was really hard, but it was really valuable … [considering] to what extent they have the sorts of higher order thinking, understanding, and skills.
>
> *(Teacher educator, Touchpoint Session, 2017)*

> I think it helped me a lot. It helped me really get into the mindset. It helped me connect that theory and then actually put it into practice. So it helped me get in, like I've got to look at the data, I've got to plan, teach, assess, and evaluate … it really helped me get in that frame that when I go into the classroom I'm there about the students, and I'm going in and I'm doing all this for the students, so it really helped me get in the right mind frame.
>
> *(Teacher education graduate, Interview, 2018)*

In addition to preservice teachers' self-awareness of their emerging teacher identity, school leaders and teacher educators also described their experiences of seeing this growth. As can be heard in the segments below, they described preservice teachers as being "much more confident than I have seen in the past."

> They are much more confident in the classroom and in organizing what they need to organize. They understand the need for some of the things that we expect of them in terms of data collection and of using our contextual knowledge to inform our practice at our school. So, I've been really impressed with that they've brought with them.
>
> *(School leader, Interview, 2018)*

> The thing that I think we need to stress is this is about the students … and my teams have told me that what they have noticed in the graduates is an increased confidence in their ability and also an increased sense of teacher identity. If they come out as strong identities, and identify strongly as a teacher, they're more likely to stay in the profession. And if this provides us with nothing else, that is the starting point from where we begin from here.
>
> *(Dean, Conference Presentation, 2019)*

Aligning the scope and sequence of teacher education

Teacher educators' talk increasingly focused on 'backward mapping' or what some referred to as 'retrofitting' or 'plugging the holes'. These were talked about as necessary to ensure the core knowledge, skills, and capabilities represented in the GTPA were taught and practiced throughout the program. Within the Collective, this was taken to mean starting from the GTPA and moving back into the scope and sequence of learning developed over the course of the program. They reported looking for the capabilities that were to be demonstrated and assessed, relating these to how and when they were taught and assessed in the program. As a group, teacher educators shared accounts of their plans for rigorous reviews of their programs: "At the moment it's more … band aid stuff but we have flagged a more rigorous kind of backward mapping later in the year" (Teacher educator, Touchpoint Session, 2017) and "We're going to implement it in the second session and then we're going to look at backward mapping through our courses" (Teacher educator, Touchpoint Session, 2017).

Teacher educators also reflected on their program satisfaction, making audible the wish that they "could wipe the slate clean and start again". In the segment below, we hear the preference for moving away from 'retrofitting' and finding 'the holes' to front-ending the assessment to inform the conceptualization and design of the program from the beginning. This reflects the growing understanding of the assessment, not as a test to be taught, but rather as a repertoire of essential professional practices and capabilities that characterize readiness for practice and that need time to develop.

> The retrofitting, backward mapping is actually quite a challenging part of the process. I just wish I could wipe the slate clean and start again. [It would] make it a lot easier if I could front-end it … trying to find the holes and the variations and the advantages and disadvantages that different courses are providing for students [preservice teachers] … So that's been the interesting part of the issue for us, finding out we have this hole in that course, but in the other course that's well and truly covered, but we've got this hole. And then [that's] complicated by how that then gets interpreted at a local level across multiple states [and campuses].
>
> *(Teacher educator, Touchpoint Session, 2017)*

In the segment above, there is a reported response to a call to action that teacher educators themselves initiated. Consider, for example, the references to "trying to find the holes and variations and the advantages and disadvantages that different courses are providing for students" and "finding out we have this hole in that course". The use of active verbs (e.g., trying, finding) illustrates how teacher educators talked of opening opportunities for critical review and evaluation of program design. They also talked of trying on new ideas: new ways of talking and new identities. This occurred not only within their own campus groups but within the larger collective of teacher educators. Throughout the process, there was risk-taking as they made the familiar strange, and the unfamiliar routine.

Despite the challenges experienced by some teacher educators in their exploits of implementation and course review, what can be heard across the talk is a significant shift in teacher educators' understandings of the GTPA. The conversations no longer centered around the new competence assessment being "tacked on" to the end of a program as "just another assessment". Instead, teacher educators began to look across their programs and investigate where, how, and the extent to which specific knowledges and skills were being developed.

The third response: Embedding the new assessment to enhance practice

The third response was an extension of the shift discussed above and in this, teacher educators and faculty leaders identify and discuss their actions in embedding the GTPA within and across programs as part of curriculum review: "In order to do the GTPA authentically and honorably, you have to build it in to the entire program. It's not tacked on the end" (Dean, Interview, 2021). The talk moves away from 'tacking on' the GTPA or absorbing it into already existing assessments, as was heard in the preceding responses. Instead, the attention turns to program cohesion as teacher educators can be heard taking up developmental perspectives on building preservice teachers' experience, knowledge, skills, and capability over the life of the program. This shift reflects the growing understanding that TPAs are expected to function as a culminating assessment designed to demonstrate classroom readiness for a beginning teacher. Teacher educators talked about making explicit provision for opportunities to be "built into the program" to ensure preservice teachers have the opportunity to learn and demonstrate the core practices for teaching: *Planning*, *Teaching*, *Assessing*, *Reflecting*, and *Appraising*. The turn to explicit teaching of using evidence to inform practice was also new for most and involved different ways of working and talking about pedagogical decision-making. New topics were introduced in programs: What counts as valuable evidence of learning? What is involved in selecting and using data? What is involved in moderation and planning using data? What does reflection look like? I know what I have taught but will I know what preservice teachers have learned and how well? A senior staff member commented on this shift to put learning at the center of attention as follows:

> The GTPA informs the content across the entire program. So, in semester one we talk about learners and development, and as part of that we will refer to elements of the GTPA and how you need to understand how learners learn effectively, what's the evidence, and how do you know what that looks like. So, we use the language pretty much from day one and talk about the GTPA.
>
> *(Dean, Interview, 2021)*

In this talk, there are references to how language use in teacher education pedagogy was undergoing change. This points to the impact of the new assessment on not only what teacher education students learn – the curriculum and pedagogical content of courses – but on their formation as a teacher and emerging professional disposition. The discussions have moved beyond those of compliance (i.e., box-ticking) and backward mapping and retrofitting, as heard in first and second responses.

In this third response, teacher educators identified the GTPA implementation as an opportunity to review their programs. In the words of one Dean, the GTPA has "given us the opportunity to revamp our programs". Teacher educators described the GTPA as providing clarity of areas for improvement, both in their own teaching and in their programs.

> When you're moderating the GTPA, it becomes clear which skills students need to work on and where their strengths lie. So, it gives you an opportunity to really review the programs and to build in the skills throughout the whole program.
>
> *(Teacher educator, Conference Presentation, 2019)*

> The effect of it really is that it's prompting program reviews in relation to the skills that students now need in schools when they're teachers in schools. For our programs, one of the clear examples was that we don't explicitly have enough teaching of literacy and numeracy teaching skills for preparing secondary teachers, so that's now something we're going to build in through each year of those secondary programs. Those are core skills that are needed. Data interpretation; being able to interpret different types of data, for example, so we're building that into different programs, making sure that it's explicitly there.
>
> *(Teacher educator, Conference Presentation, 2019)*

The talk segments below also offer insights into areas that some teacher educators identified for improvement. These included specific areas for gathering evidence (e.g., "I don't think they're addressing the integration of literacy, numeracy and general capabilities"), as well as a focus on shifting preservice teachers' mindsets about core concepts such as understanding of "planning as a process" (Conference Presentation, 2019).

> I don't think they're particularly strong in the area of feedback and I think I need to address that more deeply on the way through. I certainly do address it, but I think I need to spend more time right from second year because they're weak at providing students with effective feedback… [And that's shown] just through the marking of the GTPAs. I've been slowly embedding more around feedback in my lectures and tutorials.
>
> *(Teacher educator, Interview, 2020)*

> We have found that [preservice teachers] tend to conceptualize plans as products rather than planning as a process so it's something I am really trying to shift in their thinking – this notion that planning is an ongoing, iterative process not a static product, not a unit plan, not a lesson plan, but a process of planning and I see that mindshift occur after the GTPA. It's something that really clicks in our students once they have had that experience… So that's something that we are looking to emphasize as well as looking to draw in some more notions around adaptive expertise, what that means for improving practice.
>
> *(Teacher educator, Touchpoint Session, 2019)*

Collaboration among teacher educators was also a prized feature of GTPA implementation. In the first talk segment above, the teacher educator talked of their practice and individual action. In the second, there is an increasing emphasis on individual and collaborative action in the reference to next-step teaching plans. Improving teacher education became increasingly talked about as shared accountability.

For preservice teachers, talk also concentrated on preparation to start work in their own classrooms. Preservice teachers emphasized the value of the GTPA in teaching them "how to look for effective data" and tailoring teaching to the needs of their students. The importance of meeting the students' collective and individual needs is characterized in the following talk segments from recent teacher education graduates:

> The GTPA has helped me look at data and implement that more effectively in the classroom, and it's helped me to learn how to meet the students in the class on specific needs … while still teaching 30 kids effectively.
>
> *(Preservice Teacher, Conference Presentation, 2019)*

> One thing with the GTPA that really helps me now that I'm a teacher is not just reflecting after you have taught, but always reflecting during teaching. So, taking that data and reflecting and thinking … "What am I going to do with

> this?" and also drawing upon research to inform those decisions, and then reflecting after your formative assessment … and again through research … just a continuous cycle of practice and reflecting, practice and reflecting. And that's something that I do all the time as a teacher.
>
> *(Recent Teacher Education Graduate, Interview, 2018)*

As can be heard in the talk segment above, recent teaching graduates talked about how the GTPA has influenced their teaching practice on entry to the workforce. In the words of one graduate: "I still use that knowledge every single day when I'm in the classroom". Another stated:

> I've been really fortunate to land a beautiful job at a local school … it's a permanent ongoing position, coming out straight away as a graduate teacher … and I think the GTPA has really helped me to do that.
>
> *(Graduate teacher, Future of Education Forum, 2019)*

While the above section sheds some light on a positive turn to the GTPA, reform initiatives have repeatedly shown the need for building teacher educator confidence and wider public confidence, if culture change is to occur and be sustainable. This is discussed further below.

The fourth response: Bringing judgment and moderation into sharp focus

In this response, we hear teacher educators and education leaders talking about practices introduced to strengthen judgment consistency and fairness (see Chapters 5 and 10 for a discussion on fairness and legal implications of TPAs). The repetition of the term consistency in the segments below points to the value given to it. The centerpieces for improving consistency were calibration and moderation. The GTPA provided an opening for collaborative and evidence-informed approaches to moderation within and across institutions. In the words of an Executive Dean, the GTPA has "brought moderation into sharp focus" while also bringing to bear "cross-institutional matters" and "the need to work collaboratively" within and across participating universities. The GTPA also initiated greater recognition of the function of a common standard to give confidence in teacher preparedness on entry to the workforce. As highlighted in the words of the principal below, "teachers will be more ready than they have been in the past".

> For me what it does is it gives me confidence that we're going to have a level of consistency across the various universities, which means that there'll be a particular level of rigor that has to occur when universities are working with their graduate teachers as well as when schools are working with those same teachers. So, it does fill me with a greater sense of confidence that teachers will be more ready than they have been in the past.
>
> *(Principal, Interview, 2018)*

> I think it's important for moderation purposes and consistency across schools and across districts … there's a need to create more consistency with how we're making judgments about teacher performance. … It is important because that consistency and fair approach needs to happen … across the board.
>
> *(Mentor Teacher, Interview, 2018)*

Since joining the Collective, teacher educators identified that they had progressively implemented and strengthened calibration training within- and across-program moderation (see Chapter 7). Teacher educators talked about leading 'within course' calibration training sessions. For instance, some described holding in-person or online group sessions with staff across campuses (e.g., "we just did it all together in the one room, video-linked across three campuses", Interview, 2020). Others designed an online calibration training activity, referred to in the segment below. Readers are invited to see Chapter 7 regarding the use of decision aids, including cognitive commentaries in moderation activities.

> The calibration activity … is online. It involves modelling how to mark three samples with [accompanying] cognitive commentaries … and then a fourth mystery box sample to which they then have to submit results via a quiz form online. I get those results, what they're saying, what their marks were and what their thoughts were, and then I have a look … and I send them the cognitive commentary on that fourth sample with [the validated score] …
>
> *(Teacher educator, Interview, 2020)*

The above segment illustrates how, with the introduction of the GTPA, moderation activities were strengthened to support local quality assurance systems and processes. While some described instances of having a "very small marking team" who undertook smaller-scale moderation of predominantly borderline samples, others described intensive periods of moderation whereby larger teams of teacher educators and experienced teachers came together to score and moderate. Some reported this occurring in person and others relied on digital platforms. While digital resources have been in place since the beginning of the GTPA project in 2016, the use of digital platforms for calibration and moderation processes has become essential in the period of COVID-19. Examples of the moderation events are characterized in the talk segments below to show how they function in context.

> We have a very small marking team, very small, and one of them is me. What we tend to do [are] the Meets …[and] then when we find ones that are on the threshold, they are the ones that we are moderating … [and] will sit and work through together.
>
> *(Teacher educator, Touchpoint Session, 2019)*

> Last year we had 267 papers marked in four days. We joined up our campuses in a video conference room, so it's like one big room, and we stayed there together for that time … We do some training in the morning and

> then we mark papers. If anybody is close to the Threshold or not sure, they just turn to the person beside them, or if they don't think that it has met the threshold, they will give it back to me [Chief moderator]. I then give it to another marker … they don't know it's on threshold, I just keep track of that, and then when I get a couple of versions, if they are pretty consistent then … we come together, and we talk about it and talk through what needs to be done.
>
> *(Teacher educator, Touchpoint Session, 2019)*

A collaborative approach to moderation as a feature of small-scale and large-scale moderation at the local level is evident in these segments. We hear references to how scorers "will sit and work through together" to make decisions on threshold samples. There are echoes of this in the second segment: "we come together, and we talk about it and talk through what needs to be done". Collaboration appears to be intrinsic to how internal moderation is conducted. Value is ascribed to consulting a colleague where samples are *at* or *near* the threshold. In the small marking team, we hear "when we find ones that are on the threshold, they are the ones we are moderating", referring to peer review. In the large moderation event, we hear about training preceding marking and the preparedness of scorers to consult a colleague for cross-checking their judgment: "If anybody is close to the threshold or not sure, they just turn to the person beside them…".

In both moderation group settings, judgment practice is deprivatized in order to confirm inter-rater consistency. Underlying this is a clear appreciation of shared understandings of the standard and how it is applied within and across programs. In this approach to locally enacted moderation, judgment becomes a shared practice. It builds teacher educator confidence and confirms their identity as a recognized GTPA assessor who contributes to the range of samples that are sent forward for cross-institutional moderation online (CIM-Online™[2]; see Chapter 7). It is through the quality assurance system and processes of CIM that local judgments are scrutinized by multiple scorers from the Collective in an approach to national benchmarking. The contribution of scorers to CIM is essential for endorsement purposes and building confidence in the reliability of judgments.

As the project scaled up, the issue of workload was a recurring thread throughout teacher educators' and education leaders' talk about moderation. Many teacher educators described their own personal challenges of being time-poor with limited workload allocation for GTPA activities. One teacher educator explained: "we're constantly having to fight to make sure we have time to do the calibration and the moderation properly". In recognizing the imperative of undertaking a rigorous assessment and moderation process, teacher educators called for greater recognition of the time investment required to assess GTPA samples with integrity. This stance is characterized in the talk segment below.

> I think the moderation process is crucial … the workload for assessing GTPAs needs to be acknowledged more, rather than less, because it's that moderation

> process that takes a huge chunk of time. This is not a 1500-word essay or 1000-word essay. This is a significant [assessment] that has multiple levels all the way through, and I think the workload needs to really acknowledge that.
>
> *(Dean, Interview, 2020)*

A related concern reported by Deans was the required investment in professional development to ensure staff were sufficiently equipped to teach and assess the GTPA. The issues of staff training were tied to the increased casualization of the Australian teacher educator workforce, with many universities reporting their reliance on sessional lecturers and tutors to carry the load. As is indicated in the talk segment below, the increased reliance on a casualized workforce (1) results in increased workloads for those staff responsible for training incoming lecturers and tutors, and (2) impacts the carry-forward of knowledge and experience gained through implementing the GTPA over time.

> There is a lot of work around professional development … [and] the reliance on sessional staff means you've got a bit of a churn in staff, so you don't have the … consistency and therefore you've got a staff training problem because the GTPA actually requires a fair bit of background knowledge and so you've got that constant staff training that's needed.
>
> *(Dean, Interview, 2021)*

Intertwined in these talk segments is recognition of the significance of calibration and moderation processes if we are to get TPAs right. Regardless of workload issues, and in the absence of additional funding to support these processes, universities have committed to ensuring that moderation is conducted with integrity.

The fifth response: Realizing the potential of TPAs

In this response, the talk has shifted to a commentary on the nature of the competence assessment, its utility to improve the effectiveness of practice, and the move toward data-informed decision-making. Three purposes for using the GTPA have been distilled from the talk:

Purpose 1: Building confidence through professional learning.
Purpose 2: Connecting standards and evidence: Informing accreditation and submission of documentation to relevant state regulatory authorities.
Purpose 3: Connecting standards and evidence: Informing curriculum review and program renewal.

Building confidence through professional learning

Teacher educators reported that prior to joining the GTPA, they had limited experience of cross-course or cross-campus moderation, and little or no experience of CIM. They identified that the assessment provided new opportunities to have conversations with colleagues about evidence, decision-making, and consistency of judgments. In the words of one teacher educator, "being able to talk about what people saw as evidence and being able to come to a consensus … was where we really learned about … applying the standard" (Teacher educator, Touchpoint Session, 2019).

Teacher educators talked about CIM as benchmarking and how it built confidence in their judgment and decision-making. Supporting teacher educators to become more confident in their assessment practices was particularly relevant for more recent members of the Collective. The segment below reveals feelings of vulnerability on "heading into moderation for the first time". It also suggests a sense of appreciation and even excitement – "we're actually really excited" – about the experience of discussing CIM results (see Chapter 8 for a discussion of GTPA reports). The reports representing the results have been consistently characterized as offering "real professional learning" for teacher educator teams. While there is an official expectation that samples submitted for CIM have been moderated locally, the talk suggests that external review processes are regarded as giving local judgments legitimacy. In addition to discussions with colleagues, teacher educators look to benchmarking reports as showing the application of a common standard. The reference point for confirming judgment therefore is not one colleague but rather evidence in the report showing endorsement of judgments by the Collective. These review and quality assurance processes provide feedback on professional judgments. The feelings of excitement and nervousness surface in the first segment. In the second segment, we hear CIM associated with "consistency across institutions"; the word 'critical' is repeated, highlighting the speaker's association of CIM with both intra- and inter-institutional moderation.

> We're heading into moderation [CIM] for the first time, so we haven't received anything back … But we've talked as a team, and we just feel as though it's going to be such a supportive component of the GTPA … because going into it for the first time, you're not really too sure on whether you're making accurate assessments of the data that's been submitted; so we're actually really excited and really feel very grateful to have the opportunity to get our marking and our assessment cross-moderated. We think that that's going to be real professional learning for us as a team.
>
> *(Teacher educator, Touchpoint Session, 2020)*

> I think [CIM is] a critical component of ensuring the consistency across the institutions, and it's a very powerful tool that you can use back in the school to support professional learning and development in relation to the assessment that's going on within the institution's context, so I think it's a critical part of the process.
>
> *(Teacher educator, GTPA Touchpoint Session, 2021)*

Of particular significance was feedback provided in the reports about evidence of comparability of judgments across the Collective. The reports act indirectly as a further round of calibration; where judgments are endorsed by the Collective, the identity of the teacher educator as a reliable assessor of GTPAs is affirmed. For example, one teacher educator stated that "the results showed us that we were doing something right and [our Dean] took that feedback onboard" (Teacher educator, Touchpoint Session, 2019). Across the Collective, the growing appetite for engaging with data from CIM was clear. Teacher educators enthusiastically returned to GTPA reports as feedback on judgment consistency and program quality. More fundamentally, a new value was given to the formative potential of the GTPA and the utility of the reports for curriculum review and program renewal, as discussed below.

Informing accreditation processes

Teacher educators described their thoughts about the utility of CIM data for preparing program reports and submissions to teacher education regulatory authorities. As discussed in Chapter 8, GTPA reports, provided confidentially to each participating university, are an integral part of the feedback loop. A Dean identified how staff had used evidence from the GTPA reports for informing annual reporting. They described using aspects of the report in preparing accreditation documents required by the state regulatory authority. They characterized the reports as "of value … to demonstrate the whole story" (Dean, Interview, 2021). Here the data generated from CIM is understood as now enabling them to tell "the whole story" of their programs. Rather than serving a managerial purpose, embedded in quality and accountability agendas (Sachs, 2001), the talk segments below reveal the participants' views that the data provides "valuable insight" into program effectiveness. There is reference to telling "the whole story" in segment one, and to seeing "warts and all" in segment two. Both segments suggest an openness to share evidence of program strength and areas for improvement.

> We use it [the GTPA report] for annual reporting, and we have decided that we will provide [the regulatory authority] with the report… Because we think it's of value … to demonstrate the *whole story*. So that's been our decision … we think it's a *valuable insight.*
>
> *(Dean, Interview, 2021)*

> I think one of the things we've noticed is each year reports have grown in sophistication … also the maturity of some of the data that's coming out from things like the GTPA. And I know that you've all been working together as a Collective on the [GTPA] reports and looking at that rich data… We've seen *warts and all* data regarding a particular program or a university's approach … with a plan [for] improvement.
>
> *(Regulatory Authority Representative, GTPA Touchpoint Session, 2021)*

Of interest is the repeated reference in the first segment to 'we', denoting collective decisions and actions. In the second segment, we hear the observation from a regulatory authority, "I know that you've all been working together as a Collective on the [GTPA] reports and looking at that rich data". Overall, the talk allows us to see democratic professionalism (Sachs, 2001) in action whereby teacher educators take control by using the data to enquire into practice, extending to what has been referred to as "any trends or any conclusions that you've drawn across the accreditation period" (Regulatory Authority Representative, GTPA Touchpoint Session, 2021). This orientation is crystal clear in the following segment: "In your accreditation reporting processes you can align all of that great rich data and use it to talk about program renewal and curriculum review" (Regulatory Authority Representative, GTPA Touchpoint Session, 2021). The turn to using data as fit-for-purpose is discussed next.

Informing curriculum review and program renewal

The talk shows that teacher educators were mindful of how the data serves the official purpose of reporting how the standard has been applied at the program level. They also identified the notion of fitness-for-purpose and audience, with one saying: "I feel like, in different instances for different audiences, you might draw on one set of the analysis or another" (Teacher educator, Touchpoint Session, 2019). Emerging in the Collective was a growing awareness of the potential of the data to also inform curriculum review and program renewal. Touchpoint Sessions had provided explicit opportunities for growing data literacy through discussions of data visualization approaches in the GTPA reports. Teacher educators had been engaged in making key decisions about the selection of graphs and tables that would be most useful for their inquiry purposes. In the two segments below, we hear first from a teacher educator who suggests that the data could be used to examine the strengths and areas for improvement in programs.

> We are writing our annual report for our [Master's] program and we could use some of this sort of data to talk about … opportunities for curriculum revision or ideas for redeveloping our program by looking [at] where there are patterns in [the data] – gaps, strengths, and weaknesses.
>
> *(Teacher educator, Touchpoint Session, 2019)*

Another teacher educator identified the usefulness of having trend data to inform "our programs going forward". The availability of GTPA data was welcomed since "We've all struggled … to provide evidence when we haven't had sufficient data over a period of time" (Teacher educator, GTPA Touchpoint Session, 2021).

The theme of improvement is also evident in the talk of an AITSL representative. He highlighted the utility of the data for better connecting theory and practice, building “the assessment capabilities of teacher educators”, and informing “the delivery of programs”.

> Teacher educators gather data and evidence from TPAs [and] then use the information to review the curriculum to better connect the theory and practice, to build the assessment capabilities of teacher educators, and inform the delivery of programs.
>
> *(AITSL Representative, GTPA Touchpoint Session, 2021)*

Conclusion

The discussion has provided insights from a wide range of participants into the GTPA as a research and evidence-informed initiative. While Response One shows that some teacher educators experienced discomfort with the introduction of the GTPA, over time this gave way to a new valuing of the evidence that the assessment produced, including to show practice and decision making in the classroom. A dean revealed this shift, stating,

> The good thing about these reforms, now they’ve been implemented, is that we can now say with our hand on our heart that we know we are producing quality graduates and that’s what the TPA has given us: excellent evidence to say that through the lens of the TPA. We can say we are producing classroom ready, quality teacher graduates right across the country.
>
> *(Dean, Conference Presentation, 2019)*

A similar position was taken up by a Regulatory Authority Representative who described the TPA as “the only critical task that can be authentically defined as a reliable and valid assessment of graduate descriptors” (GTPA Touchpoint Session, 2021). She also characterized the TPA as a type of integrating device that connects “all the standards together… to demonstrate [preservice teacher] practice, to use data and evidence, to differentiate student learning” (Conference Presentation, 2019).

The corpus of talk shows a clear turn from seeing the implementation of the GTPA as a compliance measure. This has been replaced by a growing appetite for evidence that can be used for investigating program effectiveness and graduate quality and for initiating renewal. The innovation lies in how teacher educators are using evidence from the assessment for both formative and summative purposes. Stitching these two purposes together in a principled way is central to efforts to realize the potential of TPAs. It involves teacher educators knowingly engaging in democratic professionalism (Sachs, 2001) through exchanges of knowledge and expertise. These were evident in how they navigated change and shared learning in new scoring, moderation, and associated reporting processes.

The chapter has attempted to show the maturing of responses to what the TPA involves. It portrays teacher educators as trailblazers, going ahead of policy, research, and practice in Australian teacher education. There can be no doubt that ITE practice in Australia has moved ahead of theory building. The challenge therefore is for teacher education policy to catch up to practice. The GTPA Collective is well placed to contribute to this enterprise of theory building, carrying forward the learnings from our research to date.

Notes

1 Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

2 Acknowledgment: The online model of cross-institutional moderation (CIM-Online™) was conceptualized and developed in the Institute for Learning Sciences and Teacher Education, Australian Catholic University. For a discussion of CIM-Online™, readers are advised to also see Wyatt-Smith and Adie (2021).

References

Adler, P., Kwon, S., & Heckscher, C. (2008). Perspective – professional work: The emergence of collaborative community. *Organization Science*, *19*(2), 359–376. https://doi.org/10.1287/orsc.1070.0293

Australian Institute for Teaching and School Leadership (AITSL). (2011). *Australian professional standards for teachers*. https://www.aitsl.edu.au/teach/standards

Australian Institute for Teaching and School Leadership (AITSL). (2015). *Accreditation of initial teacher education programs in Australia: Standards and procedures*. https://www.aitsl.edu.au/docs/default-source/default-document-library/accreditation-of-initial-teacher-education-programs-in-australia_jan_2019.pdf?sfvrsn=4639f33c_2

Bair, M. A. (2017). Identifying dispositions that matter: Reading for consensus using a Delphi study. *The Teacher Educator*, *52*(3), 222–234. https://doi.org/10.1080/08878730.2017.1315475

Colmer, K. (2017). Collaborative professional learning: Contributing to the growth of leadership, professional identity and professionalism. *European Early Childhood Education Research Journal*, *25*(3), 436–449. https://doi.org/10.1080/1350293X.2017.1308167

Craven, G., Beswick, K., Fleming, J., Fletcher, T., Green, M., Jensen, B., Leinonen, E., & Rickards, F. (2014). *Action now: Classroom ready teachers*. Teacher Education Ministerial Advisory Group (TEMAG). Department of Education. Australia. https://docs.education.gov.au/documents/action-now-classroom-ready-teachers-report

DeLuca, C., & Johnson, S. (2017). Developing assessment capable teachers in this age of accountability. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 121–126. http://dx.doi.org/10.1080/0969594X.2017.1297010

Donaldson, G. (2010). *Teaching Scotland's future: Report of a review of teacher education in Scotland*. https://www2.gov.scot/resource/doc/337626/0110852.pdf

Freebody, P. (2003). *Qualitative research in education: Interaction and practice*. SAGE.

Hargreaves, A., & O'Connor, M. T. (2017). *Collaborative professionalism*. WISE: Qatar Foundation. https://www.wise-qatar.org/app/uploads/2019/04/rr.12.2017_boston.pdf

Racko, G., Oborn, E., & Barrett, M. (2019). Developing collaborative professionalism: An investigation of status differentiation in academic organizations in knowledge transfer partnerships. *International Journal of Human Resource Management*, *30*(3), 457–478. https://doi.org/10.1080/09585192.2017.1281830

Sachs, J. (2001) Teacher professional identity: Competing discourses, competing outcomes. *Journal of Education Policy*, *16*(2), 149–161. https://doi.org/10.1080/02680930116819

Silverman, D. (2006). *Interpreting qualitative analysis: Methods for analyzing talk, text and interaction*. SAGE.

Wyatt-Smith, C., & Adie, L. (2021). Introducing a new model for online cross-institutional moderation. In C. Wyatt-Smith, L. Adie, & J. Nuttall (Eds.), *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration* (pp. 39–58). Springer.

Wyatt-Smith, C., Alexander, C., Fishburn, D., & McMahon, P. (2017). Standards of practice to standards of evidence: Developing assessment capable teachers. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 250–270. https://doi.org/10.1080/0969594X.2016.1228603

10

TEACHING PERFORMANCE ASSESSMENTS AND CONSIDERATIONS OF POTENTIAL LEGAL CHALLENGES

Diana Pullin and Joy Cumming

Introduction

If the broad goals of addressing the public good through ensuring the fairness and effectiveness of teaching performance assessments (TPAs) can be viewed in the ways described in the previous sections of this book, another way of viewing TPAs is through an analysis of the potential legal issues that might arise when students feel that their specific needs are not addressed in TPA implementation.[1] A perspective from the United States can offer some guidance for how legal controversies might arise in the Australian context, although it should be noted that Australia has already had a notable amount of related legal controversy involving university students (Kamvounias & Varnham, 2010).

There is experience with the implementation of TPAs for credentialing in the United States (see, for example, DeVoto et al., 2021; Gitomer et al., 2019), and the many different types of uses of tests and assessments in education (Mawdsley & Cumming, 2011; Mawdsley & Williams, 2011; Pullin, 2001, 2014b, 2015) that have resulted in many legal controversies. Some of these claims were on behalf of individuals and others were brought on behalf of an entire group of similarly affected individuals, sometimes numbering in the thousands. Consideration of the U.S. controversies can illustrate the potential for legal controversies in Australia, although some particulars of how the legal issues apply will vary between two nations with similar, but different sets of laws. These areas of consideration may also be of interest to those in other legal jurisdictions. This chapter describes the various legal issues that might arise from the use of a TPA, based upon issues that have arisen in the U.S. context. The chapter then discusses the wide range of legal claims that might occur in the Australian context.

DOI: 10.4324/9780429318504-12

Potential legal issues: The United States perspective

There are three possible types of legal claims that have arisen in the United States in disputes over the use of testing and assessment in education: Claims under the U.S. Constitution or a state constitution; claims under state or federal civil rights statutes; and claims that arise under common law traditions related to negligence and contracts and state or federal statutory provisions on business relationships and consumer protection (Pullin, 2015). A new arena of legal controversy in the United States has recently arisen due to conspiracies among over 50 educators, consultants, parents, and test administrators to corrupt the college admissions testing process, leading to criminal charges and jail sentences (Korn & Levitz, 2020; Pullin, 2022). This chapter assumes that type of behavior by dozens of parents and educators in the United States will not similarly corrupt the TPA process in Australia. There are also a number of new legal claims in the United States arising from efforts to address the effects of the current pandemic on testing and assessment, particularly in the use of online approaches.

In all U.S. legal disputes, the nature of the context and consequences for someone who feels harmed are fundamental to ascertaining the role that law might play in addressing complaints about a testing or assessment program. Second, the sources of the alleged harm are important because the law restricts its coverage in particular circumstances so that, for example, constitutional claims can only be filed against public agencies, public officials, or public employees. Third, the ways in which judges or other decision-makers analyze legal complaints in any of these areas of challenge are important to consider (Pullin, 2015).

In the U.S. context, it is too early to fully gauge the impact of legal claims of unfair treatment using a teacher performance/portfolio assessment. However, previous legal activity concerning teacher testing, teacher education candidates, and judgments about their performance, as well as broader sets of issues about education testing and assessment in a variety of contexts, can help illuminate possible claims in the Australian context (Cumming & Mawdsley, 2011; *Gulino v. Board of Educ. of the City School Dist.*, 2015; Mawdsley & Cumming, 2011; Mawdsley & Williams, 2011; Pullin, 2001, 2004, 2014a, 2015).

Fairness and potential legal issues in the United States

Fair treatment and defensible decision-making by institutions is a cornerstone of U.S. law. Fairness in U.S. law is regarded as an obligation of government to treat individuals and institutions in a way that is rational or reasonable and not arbitrary or capricious. In private institutions, fairness is an obligation to honor agreements between parties and to adhere to legal obligations regulating these relationships. One consideration that can arise in any of these contexts is whether a decision-making process is sound, including the technical quality of an assessment (Pullin, 2001, 2014b, 2015).

An increasing source of potential claims of unfair or unreasonable treatment comes from those stakeholders who claim that unfair treatment intruded on their 'purchase' of services to be educated by their teacher education program or assessed by the entity overseeing the implementation of a teacher performance assessment. All of these disputes are embedded in the context of shifting perspectives on the part of government officials and the public about how education and educators should relate to society at large. These controversies have intensified as a result of the testing and educational adaptations required by the COVID-19 pandemic (Lederman, 2021; Pullin, 2022).

For any issue of fair or reasonable treatment, any outside reviewer analyzing the fairness of a situation is often required to balance the interests of a decision-maker compared with the desires of an individual impacted by the decision (Pullin, 2015). U.S. courts generally recognize the importance of state regulation of the teaching profession and are usually inclined to support government efforts to improve educator quality (Pullin, 2001, 2004, 2014a, 2015).

Fair treatment in academic decisions

Most disputes between students and universities in the United States have been resolved through administrative hearings conducted by university employees based on institutional procedures. There are ordinarily no mechanisms for review beyond the mechanisms within a university unless a student can incur the expenses and complications of filing a case in court or unless a review can be afforded under the provisions of a civil rights law. As a result, compared with the total number of disputes with students that have arisen, there is relatively little court review of university decisions except for cases involving discrimination, as will be discussed below.

The provisions for the types of issues and procedures for resolution of disputes between a university and a student are generally set out in university publications. When judges do become involved in reviewing disputes, typically they determine that students are obligated to follow requirements set by the university when they registered, signed an agreement with the university, and paid tuition and fees (Flanders, 2007).

U.S. courts have differentiated academic decisions from disciplinary decisions concerning students. While the courts offer considerable deference to academicians and administrators in their qualitative judgments about student academic performance, less latitude is shown when universities or their administrators make decisions over disciplinary matters, such as cheating or failing to follow rules. When courts do review nonacademic, disciplinary determinations by universities about their students, a failure by a university to offer procedural fairness in decision-making may lead to greater scrutiny of university decisions. Most especially clear is the obligation of a university to follow its own rules and procedures as set out ahead of time in their publications of how decisions will be made (Flanders, 2007).

Throughout U.S. history, there has long been a strong tradition of deference by judges to the professional academic judgment of faculty and university administrators, particularly in the context of professional programs. This powerful tradition of deference to academic judgment has made it difficult for students to challenge academic decisions (Flanders, 2007) as evidenced by a recent trial court case which summarized these legal standards as they had been specified in previous U.S. Supreme Court cases:

> When judges are asked to review the substance of a genuinely academic decision… they should show great respect for the faculty's professional judgment. Plainly, they may not override it unless it is such a substantial departure from accepted academic norms as to demonstrate that the person or committee responsible did not actually exercise professional judgment… University faculties must have the widest range of discretion in making judgment as to the academic performance of students and their entitlement to promotion or graduation.
>
> *(Hajjhar-Nejad v. George Washington University, 2014, pp. 116–117)*

In that case, where a medical student challenged judgments about his clinical performance and comportment, the federal trial court judge stated that it was not the role of a judge to second guess the "academic judgment of school officials that a student does not have the necessary clinical ability to perform adequately and was making insufficient progress toward that goal" (*Hajjhar-Nejad v. George Washington University*, 2014, p. 118).

However, there are exceptions to academic deference such as situations in which a student can prove that the academic judgment was so irregular, arbitrary, or capricious that deference is not appropriate. Deference is also limited if a university did not follow its own procedures as stated in its own statements of policies and procedures (Flanders, 2007). During the 2019–2021 pandemic when colleges and universities moved to programming that was entirely online but did not reduce tuition, many lawsuits were filed by angry students. Judges were forced to determine whether or not the online course requirements were actually academic decisions that were entitled to deference or were instead driven by external factors (Lederman, 2021). It is too early to know the overall outcome of all these cases.

It is worth noting that, to the extent higher education has less credibility in contemporary society, the tradition of deference by judges may become less likely. Similarly, to the extent decisions can be seen to be made by more independent authorities, like a testing company or consortium, then academic deference by a court may be less applicable. This is the case noting that at least one commentator has asserted that U.S. courts also tend to defer to testing companies (Goldschneider, 2006). Some commentators have wondered whether the impact of the pandemic on testing will include existential questions about the entire enterprise of testing (see discussion summarized in Pullin, 2022).

Discrimination law in the United States

While there has been limited court review of university decisions overall in the United States, there is one category of disputes where external authorities are more likely to become involved. In these disputes, courts have been involved in the review of university decisions because of allegations of violations of civil rights statutes and regulations. In addition, state and federal administrative agencies have the power to investigate and resolve claims of discrimination. Future teachers who are members of protected groups can present powerful legal challenges in the United States under state or federal laws, barring discrimination against members of protected groups. This set of legal protections, referred to collectively in the United States as civil rights protections, addresses a particular legal obligation to bar discrimination on the basis of disability status, minority status, and gender (Pullin, 2014b, 2015; Rothstein, 2015). For example, civil rights laws were violated when an Asian-American, required to take a 'hands-on' certification performance assessment in a workplace, was not afforded the same manner of administration or scoring of the assessment as a white employee (*Thanongsinh v. Board of Education*, 2006).

In the United States, civil rights claims alleging discrimination on the basis of race or ethnicity trigger some of the closest legal scrutinies of testing or assessment programs, including consideration of the technical quality of the assessment (Pullin, 2001, 2014b, 2015). This most often takes into account the extent to which the *Standards on Educational and Psychological Testing* of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (AERA/APA/NCME, 2014, 1999 and earlier; hereafter referred to as *Test Standards)* are followed and can result in direct scrutiny of the assessment development process. In one recent case, a judge considered, for example, whether those who participated in the development and validation process were reasonably demographically representative of the population of future teachers who would eventually be assessed (*Gulino v. Board of Educ. of the City School Dist.*, 2015).

Another consideration in U.S. civil rights cases is whether the tasks assessed represent work components actually required on the job. In one federal court case, the judge actually mapped out the overall performance standards and then determined whether they applied to the assessment scoring standards (*Gulino v. Board of Educ. of the City School Dist.*, 2015).

In addition to requirements on the technical quality of the assessment itself, U.S. discrimination laws also address the implementation of the assessment and any discriminatory impact as a result of the assessment. These issues have most frequently involved the participation of individuals with disability conditions. Accommodations in U.S. testing are required by federal civil rights laws and some additional state laws, so long as the individual has a qualifying disability protected by the law, and if the accommodation requested is 'reasonable' (Pullin, 2014b, 2015; Rothstein, 2015).

Federal laws contain broad requirements barring discrimination on the basis of disability status for categories of disabilities spelled out in the law (*Americans with Disabilities Act of 1990*; *Rehabilitation Act of 1973*). There are also specific provisions governing testing that require:

> the examination is selected and administered so as to best ensure that, when the examination is administered to an individual with a disability that impairs

> sensory, manual, or speaking skills, the examination results accurately reflect the individual's aptitude or achievement level or whatever other factor the examination purports to measure, rather than reflecting the individual's impaired sensory, manual, or speaking skills (except where those skills are the factors that the examination purports to measure).
>
> *(28 C.F.R. § 36.309(b)(1)(i) (2014))*

One case involved a request by an individual who claimed a disability and sought to take a state teacher test as an oral exam and to use a dictionary. The court held that such an accommodation would be a fundamental alteration in the content of the test and would not assess the writing skills the test was intended to measure (*Falchenberg v. New York State Department of Education*, 2008).

In another case involving an examination to qualify for a law license, a judge ruled that any modification in scoring an exam is, by its very nature, a modification which fundamentally alters the measurement of skills or knowledge the examination is intended to test, and that such a modification or accommodation is not required under the Americans with Disabilities Act (*Florida Bd. of Bar Examiners re S.G.*, 1998).

In a case involving the licensing of physicians, an appellate court ruled that the testing entity was not required to offer an accommodation that imposes an undue hardship on its program's operation andit was only required to make a reasonable accommodation (*Powell v. National Bd. of Medical Examiners*, 2004).

Human rights

The use of decision-making in public entities that is fair, in both substance and the provision of procedures for individuals or institutions to contest government decision-making, is a critical aspect of human rights as articulated by courts in the application of the provisions of the U.S. Constitution (Pullin, 2014b, 2015). In the U.S. system of laws, human rights provisions are embedded in the U.S. Constitution and in state constitutions. Violation of human rights protections, like the right to freedom of religious practice or to free expression of ideas, is another issue that can arise in the U.S. context under state or federal constitutions or statutes. For example, claims of intrusion on religious beliefs have been lodged when a future teacher implemented curriculum content or interacted with students in ways that evaluators considered unacceptable, but the student defended on grounds of personal religious beliefs (*Hennesey v. City of Melrose*, 1999) – the student lost the case. There is an increasing power of claims of religious freedoms in the United States. In the past, judges most often deferred to professional requirements if they are well-justified as being in the public interest (Pullin, 2014b, 2015), but there have been increasing and successful efforts to assert individual religious beliefs and practices that conflict with government requirements in other arenas.

Fairness of assessment

The United States has seen a small number of court cases about the implementation of a test or assessment (Pullin, 2001, 2014b, 2015). These controversies have included scoring errors (*in re Educational Testing Service PRAXIS litigation*, 2007) and printing errors (*Ellinghaus v. Educational Testing Service*, 2016).

Technical errors in teacher assessment and testing have arisen a number of times and the *Test Standards* have been utilized to address those problems (Pullin, 2001, 2014b, 2015, 2022).

In the United States, judges and government officials resolving many different types of legal disputes have taken into account the extent to which testing and assessment programs meet professional technical standards as part of their decisions under a number of different types of legal claims (Pullin, 2001, 2014b, 2015, 2022). In one study, however, many occasions were found where courts failed to address problems associated with tests that clearly failed to meet professional standards of practice (Neal et al., 2019). One example of the impact of a court's use of the *Test Standards* is a case decided in a federal trial court where the judge placed a moratorium for over a decade on a state's use of a teacher test as a factor in determining qualifications to be hired into a teaching position. The moratorium resulted from the judge's determination that the private testing company that developed the test (now part of Pearson) violated professional standards for test development (Pullin, 2015).

Other issues on testing irregularities have also arisen in U.S. courts, including cheating accusations asserted by a testing company against test-takers in a standardized testing format (Pullin, 2015). The most obvious analog to that issue concerning the implementation of TPAs would be in situations involving what seems to be fraudulent misrepresentation in performance submissions (such as the 'staged' performance or a submission that was overly assisted by program faculty, clinical supervisors, or vendors) (*Professional Standards Commission v. Denham*, 2001). In the United States, most of the legal focus for these types of errors has been on the process and procedures the testing entity used to detect irregularities and invalidate scores.

Who may be liable?

Legal claims against individuals involved in the provision of education and in assessment and testing have been limited in several different ways in the United States. There are generally limitations on liability for government employees and officers ('qualified immunity') and not-for-profit educational institutions ('charitable immunity').

It is also clear that some powerful individuals or entities have lobbied successfully to limit the liability of testing entities. For example, in 2009, the state of Ohio implemented a statute specifically created to limit any liability in that state for teacher performance assessment entities or those working with those entities unless the actions challenged were malicious, in bad faith, or wanton or reckless (*Ohio Statutes*

§ 3319.25). These types of liability shields are increasingly popular mechanisms in many different contexts in the United States, as large and small corporations lobby legislative bodies for laws to limit legal accountability to injured parties.

Validity

One of the concerns in disputes over tests and assessments in schools, universities, professional licensing, and employment decisions is often the consideration of the validity of use of a test or assessment for decisions. Validity determinations rely upon the accumulated evidence and theory supporting the use of a test score in a particular context (AERA/APA/NCME, 2014, pp. 11–41, 225). Courts often, but not always, defer to the expertise of educators or testing experts in making these determinations (Neal et al., 2019; *Gulino v. Board of Educ. of the City School Dist.*, 2015).

Reliability

Reliability evidence can be a factor courts consider in addressing challenges to tests and assessments (AERA/APA/NCME, 2014). Major concerns have been raised by researchers about the validity and reliability of the most commonly used TPA in the United States. These concerns have focused on teacher education and public policy issues (Cochran-Smith et al., 2016) and upon lack of adherence to the *Test Standards* (Gitomer et al., 2019).

Opportunity to learn

In the United States, it has been established that a high-stakes test, as important as a secondary school graduation test, cannot be utilized unless students required to take the test had a meaningful opportunity to learn the content covered by the test. This opportunity to learn consists of both adequate advanced notice of a high-stakes testing or assessment requirement, and advanced knowledge on the part of teachers to allow them to provide appropriate exposure to curriculum and instruction to prepare their students to meet the requirement (Galluzzo, 2005; Gearhart & Osmundson, 2009; Moss et al., 2008; Pullin, 2001, 2014b, 2015).

There are three possible types of learning opportunities in a TPA: Those that are provided by the teacher education institution and the cooperating clinical school sites, those afforded through the representations of competence in the scoring rubric itself, and those provided through the preparation resources offered by the test developer or testing authority.

A concern based on an opportunity to learn might be raised on grounds that there was not sufficient opportunity for institutions to prepare students for an assessment due to such factors as lack of advanced notice or lack of information about the content to be covered (Cohen & Berlin, 2020; Knight et al., 2014; Pullin, 2014b, 2015, 2022).

It is worth noting that new types of opportunity to learn issues arise in the context of the current COVID-19 pandemic. As universities and school clinical sites have closed down in response to the virus and moved to online instruction, new challenges arise asserting that the quality of education afforded is diminished, simply as a result of being online rather than receiving in-person instruction (Lederman, 2021; *Chong v. Northeastern University*, 2020).

Consumer/commercial law or negligence law violations

Increasingly, students and families have come to regard higher education as a business transaction between individuals and institutions (Korn & Levitz, 2020; Pullin, 2015, 2022). Some of the critics of TPAs in the United States frame their critiques as consumer protection concerns (Greenblatt & O'Hara, 2015). Viewing university attendance as a consumer experience opens the door to new types of legal claims against education providers and testing entities based upon legal claims that regulate commercial relationships.

Consumer law claims have been raised directly against testing programs that have experienced errors in implementing their programs, such as scoring errors (*in re Educational Testing Service*, 2007; Pullin, 2015, 2022) or accusations of test cheating (Goldschneider, 2006).

Contract law

A major legal tool for ensuring fairness in a relationship between parties or groups in business transactions is the use of written contracts. The fundamental purpose of a contract is to ensure fair treatment between those who enter into a business relationship with each other. Even though the relationship between a university and a student may be not explicitly laid out as a contract, U.S. courts have deemed the provision of education according to a university's policies and procedures, coupled with student payment of tuition and fees, to constitute a legally enforceable contract (Melear, 2003; Pullin, 2015, 2022).

There are two potential sets of claims under contract law: against the university and clinical practice site for curriculum and clinical experiences, and against the assessment program itself. The recent COVID-19 pandemic has increased the use of these types of claims by students (Lederman, 2021). For example, when a university shut down campus and took all courses and activities online, a court ruled that students were possibly entitled to a refund of facilities fees they paid to use recreational facilities. However, they were not entitled to refunds on tuition and academic fees because the contract they signed with the university did not specify that this change in instructional method would not happen (*Chong v. Northeastern University*, 2020).

Just because there may be potential contract law claims does not mean that those types of claims will result in remedies for disgruntled students. For example, the 'contract' between an individual test-taker and the testing entity can be written in such a way that a test-taker who believes they suffered unfair treatment by a testing entity might have given away the right to complain about their treatment under the agreement they entered when they registered to take the test or assessment (Pullin, 2015). U.S. courts can examine the registration materials used by an individual to sign up for a test or assessment, find that the registration effectively caused the test-taker to waive any future contract, negligence, or consumer claims against the testing company, and find no grounds for legal redress (*Ellinghaus v. Educational Testing Service*, 2016). In another case lodged against a university, a court ruled that the contract between a private university and a medical student could give the university a contractual right to expel a student for cheating on an exam (*Chenari v. George Washington University*, 2017).

Negligence law claims

In addition to contract law claims, claims of negligence in the management of a testing or assessment program have arisen in the United States. These claims assert that the use of assessment scores led to fraudulent misrepresentations about individuals that resulted in harm to reputation, to future prospects, or caused emotional harm. Here, the focus of the legal claims rests upon the notion that there is an accepted duty of care in the testing industry and that an unreasonable failure to perform that duty resulted in financial, physical, or professional reputational harm to individuals participating in the test or assessment. There are not enough cases decided by U.S. courts to determine how clearly this legal duty has emerged, but there has been legal activity that is relevant (Pullin, 2015).

An example of how these legal claims under contract law and negligence arose in the United States involved a scoring error on a teacher licensure examination where the assessment was developed and administered by the large national non-profit Educational Testing Service (ETS). Many individuals filed lawsuits against ETS because of a scoring error. Almost immediately after many of the lawsuits were filed, ETS settled the cases out of court. However, the financial consequences for ETS were payments of millions of dollars for thousands of harmed test-takers and the lawyers they retained (*in re Educational Testing Service*, 2007; Pullin, 2015). The individuals who refused to participate in the settlement pursued litigation for several years; the attorney's fee for ETS' own legal representation was no doubt considerable.

Privacy

In the U.S. context, sets of federal and state laws have been implemented to protect the privacy of students. The federal *Family Educational Rights and Privacy Act* (FERPA, 2012) and the *Protection of Pupil Rights Amendment* (2012) and their implementing regulations are the most prominent of these requirements. The law is designed to limit the inappropriate disclosure of individual school performance or disability status information. Disclosure of private student information is limited to those who have a legitimate educational need to know the information or unless a student (age 18 or older) or parent has consented to a disclosure of the information (Pullin, 2014b, 2015). These federal privacy laws would protect both future teachers participating in a TPA and also the students they teach, whose work and images might appear in the future teacher's TPA submission. It is worth noting, however, that the law and its implementing regulations are widely thought to be in need of a major update, given current technology and the widespread use of online data and data mining techniques (Russell et al., 2019).

Some states have their own provisions governing student privacy, such as California's *Student Online Personal Information Protection Act* (2014), which seeks to prevent the internet usage data of schoolchildren (but not university students) from being utilized for commercial purposes and for data mining.

Also, in the United States, there are issues related to what might be deemed 'institutional privacy' that the U.S. Congress has limited in light of what it considers accountability imperatives to address the need of the public to know about the quality of universities and, in particular, teacher education programs. Under federal law, each teacher education institution is required to report the aggregate performance data of its students on tests and assessments and on the attainment of licensure (Pullin, 2014b, 2015). This means that there are high stakes for educational institutions in addition to the high stakes for individual students.

These issues have also come to bear on teachers in some public elementary and secondary schools. For example, journalists successfully persuaded a court to use state open records laws intended to promote transparency in government to obtain access to, and then publish, the teacher evaluation scores of every single teacher in the Los Angeles California Public Schools (*Los Angeles Unified Sch. Dist. v. Superior Ct.*, 2014).

Consideration of potential TPA legal challenges: An Australian perspective

Although, as we note, legal controversies in Australian universities have been increasing in recent times (Kamvounias & Varnham, 2010), court challenges to university academic decisions are still rare. An important difference between legal procedures in Australia and those of the United States is that adversarial action in private matters is not an underlying propensity. A strong culture of mediation and dispute resolution has developed to resolve civil law complaints, established in statute law in many states and territories for different legal jurisdictions (see, for example, *Civil Procedure Act*, 2005 (NSW), *Uniform Civil Procedure Rules*, 1999 (Qld)). Between mediation and dispute resolution and the formal courts, lie bodies such as tribunals. Tribunals are regarded as more informal, with different evidence rules from formal courts (Downes, 2004), and able to "resolve disputes fairly, informally, efficiently, quickly and cheaply" without loss of due procedural fairness or natural justice (Downes, 2004, p. 8), a cornerstone of law and civil complaints in Australia. "Evidence [in tribunals] may be received in a form which would not be permitted in accordance with the rules of evidence" in law courts (Downes, 2004, p. 4) but rights of examination of evidence and response will still apply.

When legal matters, particularly those relating to matters such as education, do escalate to challenge in an Australian court, many are settled out of court and unreported, perhaps to prevent precedents opening the 'floodgates' to similar challenges (Cumming, 2009). Overall, all these factors have led to limited case law in matters of university student appeals and grievances.

Individual rights are not established under the Australian Constitution but are protected in Australia through statute law such as federal and state anti-discrimination laws that protect individuals from discrimination on grounds such as race, gender, language background, culture, religion, and sexual orientation. Individual rights are also acknowledged through case law establishing rights to natural justice or fair dealings in administrative law. Other common law traditions such as negligence and contract law, drawing heavily on English law traditions, are also available to Australian individuals. As in the United States, the outcomes of a legal complaint will depend on the context, evaluation of evidence, and potential consequences and remedies.

Given the role a TPA plays in the certification of a graduating initial teacher education (ITE) student for potential employment, the consequences of a failing TPA grade are clearly high.[2] A student who has failed a TPA may have several avenues in law to pursue a complaint including discrimination, negligence, failure in contract or consumer law, or lack of procedural fairness.

Many of the legal challenges that have occurred in the United States reflect development and implementation by external bodies of teacher assessment, and, more generally, standardized tests, for a range of certification purposes. The U.S. *Test Standards* provide guidance on professional expectations for such assessment development and use. As we have noted, the impact of legal claims with respect to performance or portfolio is less evident.

A key element in Australian student claims against failure on a TPA is that it is an academic assessment embedded within an ITE program. Unless a student who fails a TPA can avail themselves of the statutory or common law grounds for legal challenge noted previously and discussed later, student complaints about an assessment outcome will fall within academic grievance or complaint procedures of each institution, and possibly program requirements.

Fair treatment of student appeals against academic decisions

A review of Australian university academic complaint or grievance policies identified that policies state that university decisions at all times, and with respect to student appeals or complaints, should follow principles of natural justice or procedural fairness, as noted previously. Although 'natural justice' as a term lacks clarity, 'procedural fairness' is a more specific aspect (Robertson, 2015), defined as a process yielding a "fair hearing, not a fair outcome" (para. 8), with "common law duty to act fairly, in the sense of according procedural fairness, in the making of administrative decisions which affect rights, interests and legitimate expectations, subject only to the clear manifestation of a contrary statutory intention" (para. 9, citing Mason J). For university student assessment complaints, this means information about the process to follow and opportunity to submit their case, timeliness in dealing with a complaint, and impartial and unbiased decision-making (see, for example, University of Queensland [UQ], 2019b, Student Grievance Resolution, 2 Definitions, Procedural fairness (natural justice)). An individual student dissatisfied with their TPA outcome may argue that their performance has not been assessed appropriately against the stated standards or criteria, and therefore procedural fairness has not occurred.

Overall, Australian university policies for assessment complaints regarding grade outcomes or administration indicate standard procedures, with an expectation of internal resolution. An issue should be raised first with the 'relevant' decision-maker, within a stated period of time (see, for example, Australian Catholic University, 2019), "to attempt resolution" (La Trobe University, 2016, p. 5). If not resolved satisfactorily, complaints progress through an internal hierarchy of academic administrators and committees and from informal to more formal processes. Universities may offer processes for internal mediation (see, for example, University of Tasmania [UTas], 2008) or a university-appointed but independent arbiter such as an ombudsman or external reviewer (University of Notre Dame Australia, 2016).

If, following internal procedures, the student is not satisfied with the university's decision, external appeal processes are identified in a number of university policies, such as an external ombudsman (see, for example, Federation University, 2019; UTas, 2008). Higher levels of appeal that engage with legal principles are identified in some university complaint procedure policies such as the (Queensland) Administrative Appeals Tribunal for administrative decisions (Christian Heritage College, 2020), human rights and equal opportunity or anti-discrimination commissions at state or federal levels for complaints that argue breach of human rights or discrimination (see, for example, James Cook University, 2018).

A number of university policies are somewhat silent on student access to external appeal beyond the highest university committee (see, for example, RMIT University, 2020), while others indicate the right to such appeal but "normally only… after exhausting all of the avenues of resolution available within the University" (University of Southern Queensland, 2019, p. 6). Circumstances under which administrative law may apply in student challenges include appeals with respect to decisions related to academic assessment or progress of students (Rochford, 2005). However, in order to intervene in an academic decision, the court must have appropriate jurisdiction to consider a complaint.[3]

We discuss potential challenges to assessment outcomes on the basis of fairness of the assessment, or procedural fairness, in a later section. We also address Australian legal considerations of assessment validity, and university decision-making, more generally. However, as the most common avenue of appeal by Australian university students is a claim of discrimination in the assessment, where students may appeal to a number of external tribunals or courts at state and federal levels, we turn to this area next.

Discrimination law in Australia

As in the United States, discrimination is an area where students have been most successful in challenging university decisions. Such challenges may occur under omnibus anti-discrimination legislation in each Australian state or territory[4] or specific federal anti-discrimination legislation on grounds such as race and disability.[5] Discrimination in assessment may be direct, for example, a student "denied the opportunity to participate in an assessment process because of his or her disability" or indirect, for example, "imposition of unreasonable policies and conditions that disadvantage a person because of his or her disability" (Dickson & Cumming, 2018, p. 319).

The federal *Disability Discrimination Act*, 1992 (Cth) defines disability very broadly, including physical conditions, 'learning differently' from those without the disability, emotions, and behavior that is a 'symptom' of a disability (s. 4, Interpretation, Disability). Sublegislation has been enacted to address discrimination in education, the *Disability Standards for Education*, 2005 (Cth) (DSE, 2005) ['Standards']. The terminology used in the Standards is imprecise, for example, education institutions must provide 'reasonable adjustments' for the students to be able to 'access' programs and facilities for students with disability 'on the same basis' as for students without disability. No definitive case establishing precedent to interpret these terms yet exists (Dickson, 2015). The Standards address assessment specifically and institutions may need to adjust assessment procedures and conditions "to remove barriers to a student's ability to display his or her knowledge and skills" (Dickson, 2015, p. 158). The Standards allow the education provider to maintain 'integrity' of an academic course or program:

> In providing for students with disabilities, a provider may continue to ensure the integrity of its courses or programs and assessment requirements and processes, so that those on whom it confers an award can present themselves as having the appropriate knowledge, experience and expertise implicit in the holding of that particular award.
>
> *(DSE, 2005, 3.4(3))*

However, the main expectation for university assessment is provision of reasonable adjustments to enable a student with disability to complete the assessment fairly. Universities have tended to be generous in provision of assessment adjustments for students, as long as validity, reliability, transparency, fairness, effectiveness, efficiency, and balance of impact on all stakeholders are maintained. Exemptions from assessments are not considered to be a reasonable adjustment. Dickson (2015) outlines a range of assessment adjustments that have been considered reasonable in Australian case law for students with disability, ranging from format including font, paper color and size, and paper 'masks', alternatives to modes (writing, reading, hearing, speaking), time and scheduling, and level of achievement thresholds.

Exemptions from assessments are not considered by universities (see Griffith University, 2017) or the courts (*Sklavos*[6] *v Australasian College of Dermatologists* [*Sklavos*], 2017) to be a reasonable adjustment. Further, reliance on a compulsory assessment for professional certification has been noted as "lawful" (*Sklavos*, 2017, para. 217, Bromwich J). Conversely, one case has found that excusal by the education provider on the basis of absence or illness was not discriminatory if posited in the student's interests, but not denying the student participation if they wished. This case occurred in a context where the assessment was not high-stakes, such as a TPA (Dickson & Cumming, 2018).

Therefore, to appeal a failing grade on a TPA (or a complaint that a higher grade was warranted, as has occurred in some cases), a student would need to establish that lack of a reasonable adjustment has affected the student's potentially successful performance. While there have been numerous legal challenges on the basis of discrimination, and a small number of rulings about the nature of adjustments that are reasonable, the onus on the student to establish the claim means that few have been successful (Dickson & Cumming, 2018). This may also be due to the tendency of Australian courts and tribunals, as we have noted for the United States and other countries, to defer to the professional expertise and academic independence of universities in matters of policy and academic decision-making (Farrington & Palfreyman, 2012; Kamvounias & Varnham, 2006; Lindsay, 2007; Rochford, 2015).

Australian discrimination law addresses not only disability but a range of other student characteristics that may negatively impact their achievement. A further issue related to special populations is whether there has been any adverse impact on specific student groups. In another eventually unsuccessful medical certification challenge, an overseas trained doctor (OTD) (*Australian Medical Council v Sir Ronald Wilson, Elizabeth Hastings, Jenny Morgan, Dr B Siddiqui and Commonwealth Minister of Health*, 1996 [*Siddiqui*]) alleged racial discrimination regarding the imposition of a quota on access to a certification examination.[7] To establish racial discrimination, Dr Siddiqui needed to establish that overseas trained doctors of his ethnic origin were a lower proportion of all OTDs gaining admission, and the quota requirement, following language similar to such challenges in the United States, "had a disproportionate adverse impact on OTDs of Indian national origin" (para. 68). In Australia, concerns may be raised about the impact of a TPA on preservice teachers with Aboriginal or Torres Strait Islander backgrounds and others from diverse cultures or non-English speaking backgrounds. Increased teacher and school leader diversity in Australia has been identified as essential to improve learning outcomes for all students (Australian Institute for Teaching and School Leadership [AITSL], 2019). However, public information on any adverse impact of TPA requirements on preservice teachers from different backgrounds is not available.

Fairness of the TPA

Preservice teachers who fail a TPA, and do not have a discrimination claim, may consider legal issues with respect to the fairness of the TPA assessment and process itself, similar to those of the United States. In the absence of similar standards in Australia, such claims may draw on the U.S. *Test Standards* (AERA/APA/NCME, 2014) to justify complaints. Primary issues are likely to concern the validity of the assessment for its purpose and interpretation, reliability of grading, and opportunity to learn. Such challenges may occur within a university, following the complaints procedures previously discussed, or in courts or tribunals under principles of administrative law, or the common law tort of negligence.

Who may be liable?

An important aspect in the establishment of a TPA in Australia is accreditation by AITSL, as advised by an expert group acting on behalf of the organization (AITSL, 2017a). AITSL therefore has an authority role in stating an accredited TPA is fit-for-purpose, although it does not have day-to-day responsibility for ensuring its implementation. Processes for ongoing monitoring of a TPA implementation are noted as potential, but not yet explicated, evidence requirements (AITSL, 2017a). Although AITSL has a role in endorsing TPAs across Australian universities following review and evaluation by the expert group, its website indicates that it has no liability for any errors, loss, or damage related to website material, which includes information on accredited TPAs and accreditation processes (AITSL, 2017c). Liability for any claims against TPA fairness may therefore rest with the developing body but may also relate to processes of implementation.

Validity

Accreditation processes require evidence of validity. Australian cases in employment law have examined the validity of psychometric tests used for employment selection and considered job-related validity in weighing evidence (Cumming & Mawdsley, 2011). However, given the procedures outlined in the accreditation of TPAs in Australia, and emphasis on such assessments aligning with professional standards and conceptions of classroom readiness, an individual student would be unlikely to be able to appeal successfully against the validity of a TPA as an appropriate assessment of their achievement against meeting the professional standards and the notion of 'classroom readiness'. A student may be able to argue that an institution has been negligent in its implementation of the accredited TPA, which is discussed in a later section.

Reliability

As noted, a critical component of an accredited TPA is "robust processes" (AITSL, 2017b, p. 12) for 'moderation' of student outcomes, within and across institutions, "to ensure consistent scoring between assessors, and consistent decision-making against the achievement criteria, including to separate those that meet the standard and those that do not" (p. 19). The Graduate Teacher Performance Assessment (GTPA®)[8] is a complex performance task integrating evidence from a range of sources with the rubric for assessment drawing on multiple criteria reflecting the AITSL expectations. It results in a single judgment as to whether a student meets the necessary standard. Although universities may score and use GTPA outcomes to contribute to overall Grade Point Average (GPA) calculations, the primary purpose for the GTPA, and work on validation and reliability that has been undertaken, is designation that the graduating teacher is classroom ready.

The justification that processes of grading and moderation meet accreditation expectations, while expressed vaguely, is therefore important. Grading should be undertaken by "well-trained assessors" in the application of the TPA rating scale or rubric (AITSL, 2017b, p. 12). University academic appeal policies can indicate that a student who is dissatisfied with a grade may request a re-mark of assessment by an independent marker against the assessment criteria. One university policy states the independent marker should "where possible… be provided with examples of different levels of performance against the criteria and standards" (UQ, 2019a, p. 11). Such guidelines do not usually state the expectation of the capacity of the person who undertakes the re-mark, perhaps assuming competence and knowledge equivalent to those who undertook the initial grading. However, AITSL guidelines note the reliability issue of assessors "loosely connected" to a university (AITSL, 2017b, p. 12). This could also apply to assessors who have not undergone appropriate training in the application of the TPA rubric. These concerns reflect the former U.S. *Test Standards* guidelines (AERA/APA/NCME, 1999) that critical human judgments should be undertaken by raters "well-qualified… to apply their knowledge and experience to reach meaningful and relevant judgments that accurately reflect their understandings and intentions" (p. 54), and that such individuals should be provided "adequate training and instructions" (p. 48). It is therefore important for fairness and reliability that TPAs are marked by appropriate assessors. Chapter 6 in this book has indicated the strong focus on reliability and Chapter 7, moderation processes for calibration of markers for the GTPA. For TPAs more generally, and the high-stakes nature of the assessment, an issue may be ensuring that in appeal processes, a re-mark of a challenged TPA outcome is undertaken by a marker with the training and understanding to maintain the critical standard for 'classroom readiness'.

An element of reliability that may also affect validity is differences in place in different institutions to assist students who are not immediately successful on a TPA. For the GTPA, these may include assistance with revision of core components of the GTPA assessment task, as well as opportunities to repeat. The number of times a student may repeat a TPA within a program will be governed by the university. As we noted, concerns have arisen in the United States regarding 'over assistance' by university or supervising staff, and potential fraudulent misrepresentation of a student's own capabilities on a TPA. It is not clear how students' perceptions of such differences would form a basis for a student complaint within a university, if a student considered they were disadvantaged in comparison with students in other universities.

Opportunity to learn

The recommendation to introduce a TPA to measure graduate teachers' competence and assure that they are classroom ready was published in 2014 (Craven et al., 2014). Successful completion of a TPA became a graduation requirement for ITE students in 2019. Given ITE programs are generally of four years' duration or less, the first cohort completing TPAs should have been aware of the requirements on enrollment, meeting expectations for adequate notice for opportunity to learn, as established in U.S. case law (Moss et al., 2008).

There are three possible types of learning opportunities in the GTPA: Those that are provided by the ITE institution generally, those afforded through the representations of competence in the scoring rubric itself, and those provided through the GTPA preparation resources offered by the Institute for Learning Sciences and Teacher Education (ILSTE) at Australian Catholic University (ACU). In the GTPA, the opportunity to learn issue could arise for any future teacher in terms of whether or not the ITE program provided a meaningful opportunity to learn the knowledge, capacities, and professional performance skills needed to succeed on the GTPA. This claim focuses on the ITE provider most directly. However, as noted, given the processes for accreditation of both ITE programs and TPAs, the expectation is that there is a high degree of validity and match between program content and the TPA assessment. During periods of trialing, as part of the ongoing validation and implementation of the GTPA, universities identified areas of programs that needed strengthening to ensure students were prepared to undertake their GTPA assessment. Hence, prior to implementation of the GTPA, there was a 'washback' from the assessment, matching AITSL's Program Standard 1.2 accreditation expectations, on ITE content.

Responsibilities for ensuring adequate and meaningful opportunity to learn what a teacher needs to know and be able to do, as measured by a TPA, go beyond the university program. Schools that provide placement sites for student fieldwork are also critical in such opportunities. For the GTPA, guidelines have been established for teacher educators to liaise and collaborate with schools and supervising teachers "to ensure all participants are informed about the purpose and implementation of the GTPA during the professional experience placement" (ILSTE, ACU, 2020). Minimum expectations for ITE student professional experience are outlined in accreditation procedures for ITE programs, including roles and responsibilities of the key groups that share responsibility for the professional experiences, with resources and quality case studies available (AITSL, 2015). Overall, given intended ITE program content, professional experience expectations, and alignment, students seeking to challenge TPA failure would therefore have to meet the onus established under a claim of negligence, that is, that the enacted program content, did not match the intended.

Consumer/contract law and negligence

Consumer/contract law

As in the United States, universities position themselves as businesses and students as clients (Kamvounias, 2015). Inevitably, students therefore see themselves as consumers engaged in a contract with a university, implicitly based on enrollment terms

(Rochford, 2015). However, Australian legal precedent for challenges in contract law for university students is limited, due to the availability of alternative grounds such as discrimination and administrative law (Rochford, 2015). A claim under contract law by a student for a failed TPA would need to establish the enrollment contract, and payment of any fees, warranted a contract to educate and, in corollary, the university's "failure to teach" the appropriate curriculum for the student to succeed (p. 88). While such challenges are reported in the media from time to time, matters have been resolved by unreported settlement, or not pursued. As in many common law areas in Australia, provision of services under $40,000, which would apply to a number of university programs, is governed by federal Australian consumer law under the *Competition and Consumer Act*, 2010 *(Cth)*. Provision of services such as university teaching must be "rendered with due care and skill" (Corones, 2012, p. 9), requiring proper qualifications of staff who deliver the course. The extent to which claims are made about "specialist skills and expertise", requires a higher expectation of provision of service "than in normal circumstances" (p. 10). Given these expectations, a dissatisfied student may consider challenging a university for negligence, rather than in contract law.

Negligence law claims

Negligence is a common law tort that may be available to students. To establish negligence requires a duty of care by the university, loss (failure to graduate due to failure on a TPA) "sustained by a student as a consequence of the institutions' or the teacher's failure to educate at an appropriate standard" (Horton et al., 2015, p. 186), causation, and identification of an appropriate remedy. Unless a student could establish that the grading of their individual TPA lacked professional care, a negligence claim would require evidence that a group of students had not been successful due to such failure (Cumming & Mawdsley, 2011). As of 2016, no educational negligence case had been reported as successful in Australia (Cohen, 2016) and whether a university may be liable for educational negligence or 'malpractice' is not resolved in Australia (Horton et al., 2015). Several Australian states have established civil liability acts, such as the *Civil Liability Act*, 2003 (Qld), to replace common law negligence claims, with primary principles that "the risk of harm", as a result of breach of duty, was "foreseeable", "not insignificant", and "a reasonable person" would have taken precautions (s. 9). The onus is on the plaintiff to establish the breach and impact.

There are reported instances of failure to teach to an appropriate standard in Australian education. In one example, students in a secondary school in New South Wales (NSW), whose English results were in the lowest 20%, while results for other subjects for the same students were in the top 20%, alleged negligence but settled out of court (Williams, 1996). In another case in NSW, it was identified that students were studying the incorrect mathematics syllabus for their subject and would have two months to 'catch up' for the examination. Various statements were made about the support the students would be given and options available to them (O'Connor, 2017). No information on outcomes for these students or any legal challenges has been identified. In such cases, an issue is not just whether a wrong occurred, but also to identify the impact and loss that were a result and an appropriate remedy. Such arguments would usually be based on loss of income from a future program of study or, in the case of university students, employment, although, for the students, certainty of employment could not be assumed. In some circumstances, "loss of chance" may be sufficient for the court to identify potential loss of income and damages (Rochford, 2001, p. 327). However, a further element in claims alleging failure to learn under contract or the tort of negligence would be to establish causality, that is, that the actions of the institution were responsible for a failure to learn. As has been noted in various court decisions, failure to learn may be the result of a lack of engagement by the student or personal factors such as illness or family matters (Rochford, 2001).

Negligence for a TPA could extend to processes for establishing a TPA, establishing a cut-score, and grading of the student work. However, as noted, the current processes for developing a TPA and its accreditation reduce the possibilities that such a claim would be successful.

Privacy

Australians do not have absolute privacy protection. However, individual privacy is protected by state and federal privacy acts (see, for example, the *Privacy and Personal Information Protection Act*, 1998 (NSW), the *Privacy Act*, 1988 (Cth)), usually "balanced with the interests of entities in carrying out their functions or activities" (*Privacy Act*, 1988 (Cth), s. 2(b)). Because of the nature of their establishment, Australian universities are "public authorities" or "government agencies" governed by such federal, state, and territory legislation (Fleming, 2015, p. 65). Teacher registration authorities such as the Queensland College of Teachers are also governed by relevant privacy legislation such as the *Information Privacy Act*, 2009 (Qld). A number of privacy principles inform the treatment of individual information in Australia more generally. A core principle relates to the need for individuals to consent to the provision of data from one entity to another. Processes for provision of student assessment information may differ across Australian accreditation authorities. However, in many circumstances, the institution may provide the registration authority with the necessary information for new graduates (see, for example, Queensland College of Teachers, n.d.;Victorian Institute of Teaching, 2020). Such transfer of student assessment information would require student consent. Individual TPA results would therefore constitute part of such information. At the time of writing, publication of student outcomes, or overall success rates, by institutions on a TPA is not planned.

An Australian caveat

In legal cases, different standards of proof apply to decision-making, based on the available evidence, in criminal law and civil law. In criminal law, the well-known standard is 'beyond reasonable doubt'. In civil law, simply stated, the standard is 'on the balance of probabilities'. However, in Australian law, an English consideration of the civil law standard has been implemented, known as the *Briginshaw* (*Briginshaw v Briginshaw*, 1938) standard. The Briginshaw standard remains the balance of probability but incorporates an extended understanding of the concept of probability – that the more serious the alleged (mis)conduct, the less the probability that the event would have occurred. Hence the more serious the alleged action, the greater the need for cogent evidence, with "rigour… and objective analysis" in collection and consideration of cogent evidence (*White v State of Queensland [Central Queensland Hospital and Health Service]*, 2017, p. 17 or para. 54). The principle is incorporated in civil proceedings in the Australian *Evidence Act*, 1995 (Cth) where a decision on the balance of probabilities may be influenced by the "nature" of the case and the "gravity" of the allegations (p. 105, s 140). The Briginshaw principle has been used in law when a person's employment or livelihood is at stake, including cases about unfair dismissal claims, with a higher onus on an employer to establish that an employee should be dismissed (Cumming & Mawdsley, 2011). More specifically, it has been applied in decisions regarding teacher registration and employment, often in cases of allegations of teacher misconduct, both with respect to the serious consequences of inadequate investigation but also in terms of the onus of proof to prove a person is no longer suitable to teach (see, for example, *Queensland College of Teachers v DYR*, 2016 ['*DYR*'][9]). In other circumstances, the onus may be on the employee to demonstrate suitability to teach (*Queensland College of Teachers v Teacher CXJ* (No 2), 2017).

The Briginshaw 'standard of proof' has been adopted by all Australian anti-discrimination jurisdictions as a rule based on the general belief that any allegation of discrimination or harassment is a 'serious matter', although how and when it should be applied has been questioned (de Plevitz, 2003). For a student making a legal challenge with respect to an unsuccessful TPA outcome, and hence deprivation of employment as a teacher, the onus may require an institution or TPA developer to provide the cogent evidence that the assessment is an appropriate (valid and reliable) indicator of the likelihood the student is or is not 'classroom ready' to teach. This may well be an area where future research is needed to examine the predictive validity of a TPA for a graduate's classroom readiness to teach and have positive effects on student learning outcomes. For a discrimination claim by a student, the onus is on the student to establish to the Briginshaw standard that discrimination has occurred.

Conclusion and recommendations

In the United States, there have been decades of effort to improve teacher education; wave after wave of reforms were initiated, yet persistent concerns about teacher education and teacher quality remain (Pullin, 2017). Legal controversies in the United States have often resulted from these initiatives. For Australia, might recent efforts to reform teacher education and entry to the teaching profession have similar outcomes?

In many respects, the current Australian approach through initiatives like the GTPA presents a promising prospect. The GTPA is one component of an effort to take a systemic approach to achieve meaningful reform of the teaching profession and teacher education. Is it sufficient to effectively enhance the opportunity to learn and the goals of the reforms? Will it do so both fairly and effectively? Will it withstand potential legal challenges? The answers to these questions will depend, in large part, on the choices that have been made and continue to be made about the design and implementation of the GTPA and the other aspects of the current Australian education reform initiatives.

Both U.S. and international scholars have noted the importance of a systemic approach (Darling-Hammond et al., 2017) and an 'improvement science' approach (Bryk et al., 2016) to education reform. The success of the GTPA, as well as the other components of current reform in Australia, will depend in large part on the extent to which governments, scholars, and practitioners are determined to stay the course and commit the resources for the difficult work to be done. Included in this work must be efforts to implement ongoing discernment of potential legal challenges to these initiatives, and how, or whether, to respond. Further, ongoing research on the quality of the assessment and the consequences of the implementation of the GTPA will be essential.

Notes

1 Faculty might have concerns about a TPA and its implementation and institutions might as well. This chapter does not address those types of issues, but see Pullin (2004) for a discussion of some of these potential legal issues.

2 Since 2016, Australian teacher education students are also required to pass the Literacy and Numeracy Test for Initial Teacher Education (LANTITE) as a component of the reforms to improve confidence in the quality of teaching and classroom readiness of new teachers (AITSL, 2017d). LANTITE is therefore another high-stakes assessment for prospective teacher graduates but may be taken at any time, even prior to enrolment, and is not argued as dependent on program content, or as predictive of 'classroom readiness'.

3 Although universities may be established under state or territory legislation, and standards are overseen by a national quality assurance body, Tertiary Education Quality and Standards Agency, they are 'self-regulatory' bodies that establish their own 'academic governance' involving regulations and policies with respect to all areas of their operation (see Varnham, 2015). In a noted administrative law case, *Griffith University v Tang* (2005), following a successful challenge to a university decision to exclude the student, upheld on appeal, a further appeal by the university to the High Court led to determination that the university decision was made under policy, not under its statutory origins, was not subject to judicial review, and the court did not have jurisdiction to intervene.

4 State and Territory laws address all protected attributes: *Anti-Discrimination Act 1977* (NSW), *Equal Opportunity Act 1984a* (SA), *Equal Opportunity Act 1984b* (WA), *Anti-Discrimination Act 1991* (Qld), *Discrimination Act 1991* (ACT), *Anti-Discrimination Act 1992* (NT), *Equal Opportunity Act 1995* (Vic), *Anti-Discrimination Act 1998* (Tas). These acts address discrimination on a range of characteristics, for example, the

Anti-Discrimination Act 1991 (Qld) lists (a) sex; (b) relationship status; (c) pregnancy; (d) parental status; (e) breastfeeding; (f) age; (g) race; (h) impairment; (i) religious belief or religious activity; (j) political belief or activity; (k) trade union activity; (l) lawful sexual activity; (m) gender identity; (n) sexuality; (o) family responsibilities; (p) association with, or relation to, a person identified on the basis of any of the above attributes.

[5] Federal laws are directed to particular protected attributes, for example, the *Disability Discrimination Act 1992* (Cth), *Racial Discrimination Act 1975* (Cth), *Sex Discrimination Act 1984* (Cth), *Age Discrimination Act 2004* (Cth).

[6] A recent unsuccessful legal challenge for failure to pass such a high-stakes examination to enter a medical specialization has examined exemption from the assessment as a reasonable adjustment. The claimant indicated a phobia against assessment, that increased to the point of being unable to undertake any examination, and arguing a waiver was necessary. The professional college was willing to provide reasonable special considerations. The initial Federal Court judgment held that a "waiver" was not a "reasonable adjustment" (*Sklavos v Australasian College of Dermatologists* [*Sklavos*], 2017, para. 59).

[7] For Australian certification, he was required to pass two examinations, first a multiple-choice test, required as a prerequisite to a second case appraisal assessment. In a complicated situation, Dr Siddiqui was unsuccessful on the first test but eventually passed. However, by then his passing score was insufficient to be placed in a quota that had been introduced to restrict the numbers of OTDs allowed to proceed to the second assessment.

[8] Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU) and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

[9] A teacher charged with sexual misconduct was suspended from registration and identified as "not suitable to teach" (DYR, 2016, p. 2). Criminal charges against the teacher were not found. The teacher appealed the administrative decision in an administrative tribunal. The Tribunal reasoned that it was "not satisfied on the Briginshaw standard" that evidence was sufficient to establish that alleged behavior took place and ordered the suspension be ended (p. 5).

References

Age Discrimination Act 2004 (Cth).

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Americans with Disabilities Act of 1990, 42 U.S.C. 12101.

Anti-Discrimination Act 1977 (NSW).

Anti-Discrimination Act 1991 (Qld).

Anti-Discrimination Act 1992 (NT).

Anti-Discrimination Act 1998 (Tas).

Australian Catholic University. (2019). *Student appeals procedures*. https://handbook.acu.edu.au/1421715

Australian Institute for Teaching and School Leadership (AITSL). (2015). *Professional experience. Participant roles and responsibilities*. https://www.aitsl.edu.au/docs/default-source/default-document-library/professional-experience---participant-roles-and-responsibilities.pdf?sfvrsn=b187e33c_0

Australian Institute for Teaching and School Leadership (AITSL). (2017a). *Teaching performance assessment: Program Standard 1.2*. https://www.aitsl.edu.au/docs/default-source/initial-teacher-education-resources/tpa/tpa-fact-sheet.pdf?sfvrsn=1410cb3c_6

Australian Institute for Teaching and School Leadership (AITSL). (2017b). *Teaching performance assessment services. Principles of operation*. https://www.aitsl.edu.au/docs/default-source/initial-teacher-education-resources/eag-operational-principles.pdf?sfvrsn=b90cfd3c_12

Australian Institute for Teaching and School Leadership (AITSL). (2017c). *Terms and conditions*. https://www.aitsl.edu.au/general/terms-and-conditions

Australian Institute for Teaching and School Leadership (AITSL). (2017d). *Understand the literacy and numeracy test*. https://www.aitsl.edu.au/deliver-ite-programs/learn-about-ite-accreditation-reform/understand-the-literacy-and-numeracy-test

Australian Institute for Teaching and School Leadership (AITSL). (2019). *Spotlight. Diversity in school leadership*. https://www.aitsl.edu.au/docs/default-source/research-evidence/spotlight/spotlight-diversity-in-school-leadership.pdf?sfvrsn=93effa3c_6

Australian Medical Council v Sir Ronald Wilson, Elizabeth Hastings, Jenny Morgan, Dr B Siddiqui and Commonwealth Minister of Health [1996] FCA 1618, (17 July 1996).

Briginshaw v Briginshaw (1938) 60 CLR 336.

Bryk, A., Gomez, L., Grunow, A., & LeMahieu, P. (2016). *Learning to improve: How America's schools can get better at getting better*. Harvard Education Press.

Chenari v. George Washington University, 847 F. 3d. 740, D.C. Ct of Appeals, 2017.

Chong v. Northeastern University (2020). 494 F. Supp. 3d 24 (D. Ma. 2020).

Christian Heritage College. (2020). *Grievance policy and procedures for domestic students – academic grievances*. https://chc.edu.au/policies/student-administration/grievance-policy-and-procedures-for-domestic-students-academic-grievances/

Civil Liability Act 2003 (Qld).

Civil Procedure Act 2005 (NSW).

Cochran-Smith, M., Stern, R., Sanchez, J. G., Miller, A. F., Keefe, E. S., Fernandez, M. B., Chang, W., Cummings Carney, M., Burton, S., & Baker, M. (2016). *Holding teacher preparation accountable: A review of claims and evidence*. National Education Policy Center. http://nepc.colorado.edu/publication/teacher-prep

Cohen, C. (2016). *Australian universities' potential liability for courses that fail to deliver. Insights. Colin Biggers & Paisley Lawyers*. https://www.cbp.com.au/insights/insights/2016/december/australian-universities-potential-liability-for-c

Cohen, J., & Berlin, R. (2020). What constitutes an "opportunity to learn" in teacher preparation? *Journal of Teacher Education*, *71*(4), 434–448. https://doi.org/10.1177/0022487119879893

Competition and Consumer Act 2010 (Cth).

Corones, S. (2012). Consumer guarantees and the supply of educational services by higher education providers. *University of New South Wales Law Journal*, *35*(1), 1–30. http://classic.austlii.edu.au/au/journals/UNSWLawJl/2012/1.pdf

Craven, G., Beswick, K., Fleming, J., Fletcher, T., Green, M., Jensen, B., Leinonen, E., & Rickards, F. (2014). *Action now: Classroom ready teachers*. Teacher Education Ministerial Advisory Group. https://www.dese.gov.au/uncategorised/resources/action-now-classroom-ready-teachers-report

Cumming, J. J. (2009). Assessment challenges, the law and the future. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century. Connecting theory and practice* (pp. 157–182). Springer.

Cumming, J. J., & Mawdsley, R. D. (2011). Certification of teachers, pre-service teacher education, tests and legal issues in Australia and the United States of America (U.S.): Part B implications for Queensland and Australia. *International Journal of Law and Education*, *16*(1), 65–86. https://www.anzela.edu.au/assets/ijle_vol_16.1_-_5_cumming_and_mawdsley.pdf

Darling-Hammond, L., Burns, D., Campbell, C., Goodwin, A. L., Hammerness, K., Low, E., McIntyre, A., Sato, M., & Zeichner, K. (2017). *Empowered teachers: How high-performing systems shape teaching quality around the world.* Jossey-Bass.

de Plevitz, L. (2003). Briginshaw 'standard of proof' in anti-discrimination law: 'Pointing with a wavering finger'. *Melbourne University Law Review*, *27*(2), 308–333. http://www5.austlii.edu.au/au/journals/MelbULawRw/2003/13.html

De Voto, C., Olson, J., & Gottlieb, J. (2021). Examining diverse perspectives of edTPA policy implementation across states: The good, the bad, and the ugly. *Journal of Teacher Education*, *72*(1), 42–55. https://doi.org/10.1177/0022487120909390

Dickson, E. (2015). Disability standards for education. In S. Varnham, P. Kamvounias, & J. Squelch (Eds.), *Higher education and the law* (pp. 149–162). The Federation Press.

Dickson, E., & Cumming, J. (2018). Reasonable adjustment in assessment – The Australian experience. In K. Trimmer, R. Dixon, & Y. Findlay (Eds.), *The Palgrave handbook of education law for schools* (pp. 315–333). Palgrave Macmillan.

Disability Discrimination Act 1992 (Cth).

Disability Standards for Education 2005 (Cth).

Discrimination Act 1991 (ACT).

Downes, G. (2004). Tribunals in Australia: Their roles and responsibilities. *Australian Law Reform Commission Reform Journal*, *84*(Autumn), 7–9. http://www5.austlii.edu.au/au/journals/ALRCRefJl/2004/2.html

Ellinghaus v. Educational Testing Service, 2016 WL 8711439, E.D. N.Y., 2016.

Equal Opportunity Act 1984a (SA).

Equal Opportunity Act 1984b (WA).

Equal Opportunity Act 1995 (Vic).

Evidence Act 1995 (Cth).

Falchenberg v. New York State Department of Education, 642 F. Supp. 2d 156 (S.D.N.Y. 2008).

Family Educational Rights and Privacy Act (FERPA), 20 U.S.C. § 1232g (2012) and Implementing Regulations under FERPA, 34 C.F.R. § 99 (2011).

Farrington, D., & Palfreyman, D. (2012). *The law of higher education* (2nd Ed.). Oxford University Press.

Federation University. (2019). *Student appeal policy*. https://policy.federation.edu.au/university/student_grievance/ch01.php

Flanders, J. (2007). Academic student dismissals at public institutions of higher education: When is academic deference not an issue? *Journal of College and University Law*, *34*(1), 21–79.

Fleming, H. (2015). 'The next two decades are going to be transparency': Regulatory challenges for universities. In S. Varnham, P. Kamvounias, & J. Squelch (Eds.), *Higher education and the law* (pp. 64–79). The Federation Press.

Florida Bd. of Bar Examiners re S.G., 707 So.2d 323 (1998).

Galluzzo, G. R. (2005). Performance assessment and renewing teacher education: The possibilities of the NBPTS Standards. *Clearing House*, *78*(4), 142–145. https://doi.org/10.3200/TCHS.78.4.142-145

Gearhart, M., & Osmundson, E. (2009). Assessment portfolios as opportunities for teacher learning. *Educational Assessment*, *14*(1), 1–24. https://doi.org/10.1080/10627190902816108

Gitomer, D. H., Martínez, J. F., Battey, D., & Hyland, N. E. (2019). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal*, *58*(1), 3–31. https://doi.org/10.3102/0002831219890608

Goldschneider, A. (2006). Cheater's proof: Excessive judicial deference toward educational testing agencies may leave accused examinees no remedy to clear their names. *Brigham Young University Education and Law Journal*, *2006*(1), 97–155.

Greenblatt, D., & O'Hara, K. E. (2015). Buyer beware: Lessons learned from edTPA implementation in New York state. *Teacher Education Quarterly*, *42*(2), 57–67.

Griffith University. (2017). *Reasonable adjustments for assessment – Students with disabilities*. https://policies.griffith.edu.au/pdf/Assessment%20Reasonable%20Adjustments.pdf

Griffith University v Tang (2005) 221 CLR 99.

Gulino v. Board of Educ. of the City School Dist., 122 F. Supp. 3d 115 (S.D.N.Y. 2015).

Hajjhar-Nejad v. George Washington University, 37 F. Supp. 3d 90 (D.C., 2014).

Hennesey v. City of Melrose, 194 F. 3d 237 (1st Cir. 1999).

Horton, R., Smith, K., & Tisbury, A. (2015). The tort of negligence and higher education. In S. Varnham, P. Kamvounias, & J. Squelch (Eds.), *Higher education and the law* (pp. 178–188). The Federation Press.

in re Educational Testing Service PRAXIS Principles of Learning and Teaching: Grades 7–12 Litigation, 517 F. Supp. 2d 832 (2007, E.D. Louisiana).

Information Privacy Act 2009 (Qld).

Institute for Learning Sciences and Teacher Education, Australian Catholic University (ACU). (2020). *Stakeholders*. https://www.graduatetpa.com/stakeholders/

James Cook University. (2018). *Student complaint management policy and procedures*. https://www.jcu.edu.au/policy/student-services/student-complaint-management-policy-and-procedures

Kamvounias, P. (2015). Students and the Australian consumer law. In S. Varnham, P. Kamvounias, & J. Squelch (Eds.), *Higher education and the law* (pp. 92–103). The Federation Press.

Kamvounias, P., & Varnham, S. (2006). In-house or in court? Legal challenges to university decisions. *Education and the Law*, *18*(1), 1–17. https://doi.org/10.1080/09539960500529850

Kamvounias, P., & Varnham, S. (2010). Legal challenges to university decisions affecting students in Australian courts and tribunals. *Melbourne University Law Review*, *34*(1), 140–180. https://law.unimelb.edu.au/__data/assets/pdf_file/0005/1703588/34_1_5.pdf

Knight, S. L., Lloyd, G. M., Arbaugh, F., Gamson, D., McDonald, S. P., Nolan, J., & Whitney, A. E. (2014). Performance assessment of teaching: Implications for teacher education. *Journal of Teacher Education*, *65*(5), 372–374. https://doi.org/10.1177/0022487114550475

Korn, M., and Levitz, J. (2020). *Unacceptable: Privilege, deceit, and the making of the college admissions scandal*. Portfolio/Penguin.

La Trobe University. (2016). *Student grievance policy*. https://handbook.acu.edu.au/handbooks/handbook_2019/general_information/student_appeals_policy/policy_on_student_appeals

Lederman, D. (2021, May 6). Courts skeptical on COVID-19 tuition lawsuits. *Inside Higher Education*. https://www.insidehighered.com/news/2021/05/06/courts-view-covid-19-tuition-refund-lawsuits-skeptically

Lindsay, B. (2007). Complexity and ambiguity in university law: Negotiating the legal terrain of student challenges to university decisions. *Australia & New Zealand Journal of Law & Education*, *12*(2), 7–24. http://www.austlii.edu.au/au/journals/ANZJlLawEdu/2007/10.pdf

Los Angeles Unified Sch. Dist. v. Superior Ct., 175 Cal. Rptr. 3d 90, 306 Ed. Law Rep. 919 (Ct of Appeal, 2nd Dist., July 23, 2014).

Mawdsley, R., & Williams, P. (2011). Teacher assessment and credentialling: The role of the federal government in a state function. *Education Law Reporter*, *262*, 735–760.

Mawdsley, R. D., & Cumming, J. J. (2011). Certification of teachers, pre-service teacher education, tests and legal issues in Australia and the United States of America (U.S.): Part A Context, and U.S. history. *International Journal of Law and Education*, *16*(1), 47–63. http://kirra.austlii.edu.au/au/journals/IntJlLawEdu/2011/4.pdf

Melear, K. (2003). The contractual relationship between student and institution: Disciplinary, academic, and consumer contexts. *Journal of College and University Law*, *30*(1), 175–233.

Moss, P., Pullin, D., Gee, J., Haertel, E., & Young, L. (2008). *Assessment, equity and opportunity to learn*. Cambridge University Press.

Neal, T., Solbogin, C., Saks, M., Faigman, D., and Geisinger, K. (2019). Psychological assessments in legal contexts: Are courts keeping "junk science" out of the courtroom? *Psychological Science in the Public Interest 20*(3), 135–164.

O'Connor, L. (2017, August 30). Major error in HSC mathematics. *The Coonamble Times*. https://www.coonambletimes.com.au/major-error-hsc-mathematics/

Ohio Statutes § 3319.25

Powell v. National Bd. of Medical Examiners, C.A.2 (Conn. 2004), 364 F.3d 79, opinion corrected 511 F.3d 238.

Privacy Act 1988 (Cth).

Privacy and Personal Information Protection Act 1998 (NSW).

Professional Standards Commission v. Denham, 556 S.E.2d 920 (Ga. App, 2001).

Protection of Pupil Rights Amendment, 20 U.S.C. § 1232h (2012).

Pullin, D. (2001). Key questions in implementing teacher testing and licensing. *Journal of Law and Education*, *30*(3), 383–429.

Pullin, D. (2004). Accountability, autonomy, and academic freedom in educator preparation programs. *Journal of Teacher Education*, *55*(4), 300–312. https://doi.org/10.1177/0022487104266717

Pullin, D. (2014a). *Performance, value and accountability: Public policy goals and legal implications of the use of performance assessments in the preparation and licensing of educators. Paper commissioned by the Council of Chief State School Officers and Stanford Center for Assessment, Learning, and Equity (SCALE).* https://secure.aacte.org/apps/rl/res_get.php?fid=1504&ref=edtpa

Pullin, D. (2014b). Professional test standards in the eyes of the law. *Educational Measurement: Issues and Practice*, *33*(4), 19–21. https://doi.org/10.1111/emip.12048

Pullin, D. (2015). Performance measures for teachers and teacher education: Corporate education reform opens the door to new legal issues. *Education Policy Analysis Archives*, *23*(81), 1–32. https://doi.org/10.14507/epaa.v23.1980

Pullin, D. (2017). What counts? Who is counting? Teacher education improvement and accountability in a data-driven era. In J. Nuttall, A. Kostogriz, M. Jones, & J. Martin (Eds.), *Teacher education policy and practice: Evidence of impact, impact of evidence* (pp. 3–16). Springer.

Pullin, D. (2022). Do standards promote fairness and legitimacy in the changing marketplace for testing? In J. L. Jonson & K. F. Geisinger (Eds.), *Fairness issues and solutions in educational and psychological testing: Implications for researchers, practitioners, policy makers, and the public.* American Educational Research Association.

Queensland College of Teachers. (n.d.). *Applying for teacher registration in Queensland*. https://cdn.qct.edu.au/pdf/Applying_for_registration_fact_sheet.pdf?_ga=2.41999078.291729523.1591746789-1717174897.1591059573

Queensland College of Teachers v DYR [2016] QCAT 427 [15].

Queensland College of Teachers v Teacher CXJ (No 2) [2017] QCAT 166.

Racial Discrimination Act 1975 (Cth).

Rehabilitation Act of 1973, 29 U.S.C. 794.

RMIT University. (2020). *Assessment, academic progress and appeals regulations*. https://www.rmit.edu.au/about/governance-management/statutes-regulations/assessment-academic-progress-appeals

Robertson, A. (2015). *Natural justice or procedural fairness*. Federal Court of Australia. https://www.fedcourt.gov.au/digital-law-library/judges-speeches/speeches-former-judges/justice-robertson/robertson-j-20150904

Rochford, F. (2001). Suing the alma mater: What loss has been suffered*? Education and the Law*, *13*(4), 319–333.

Rochford, F. (2005). Claims against a university: The role of administrative review in Australia and the United Kingdom. *Education and the Law*, *17*(2), 23–41. https://doi.org/10.1080/09539960500165259

Rochford, F. (2015). The contract between the university and the student. In P. Kamvounias, S. Varnham, & J. Squelch (Eds.), *Higher education and the law* (pp. 82–91). The Federation Press.

Rothstein, L. (2015). The Americans with Disabilities Act and higher education 25 years later: An update on the history and current disability discrimination issues for higher education. *Journal of College and University Law*, *41*(3), 531–598.

Russell, N., Reidenberg, J., Martin, E., & Norton, T. (2019). Transparency and the marketplace for student data. *Virginia Journal of Law and Technology*, *22*(3), 108–159. https://www.fordham.edu/download/downloads/id/14725/transparency_and_the_marketplace_for_student_data.pdf

Sex Discrimination Act 1984 (Cth).

Sklavos v Australasian College of Dermatologists [2017] FCAFC 128.

Student Online Personal Information Protection Act, California Bus. & *Prof. Code* §§ 22584–22585 (2014).

Thanongsinh v. Board of Education, 462 F. 3d 762 (7th Cir. 2006).

Uniform Civil Procedure Rules 1999 (Qld).

University of Notre Dame Australia. (2016). *Policy: Student appeals*. https://www.notredame.edu.au/__data/assets/pdf_file/0009/2052/POLICY-Student-Appeals.pdf

University of Queensland. (2019a). *Assessment*. https://ppl.app.uq.edu.au/content/3.10.02-assessment#Procedures

University of Queensland. (2019b). *Student grievance resolution*. https://ppl.app.uq.edu.au/content/3.60.02-student-grievance-resolution

University of Southern Queensland. (2019). *Academic appeal procedures*. https://policy.usq.edu.au/documents/141633PL#4.9

University of Tasmania. (2008). *Ordinance No 8. Student complaints*. https://www.utas.edu.au/__data/assets/pdf_file/0008/76139/Ordinance-8-Student-Complaints-December-2017.pdf

Varnham, S. (2015). University governance: Responsibility and accountability. In P. Kamvounias, S. Varnham, & J. Squelch (Eds.), *Higher education and the law* (pp. 16–29). The Federation Press.

Victorian Institute of Teaching. (2020). *Registering as a graduate*. https://www.vit.vic.edu.au/registering-as-a-teacher/how-do-i-register-as-a-teacher/registering-as-a-graduate

White v State of Queensland (Central Queensland Hospital and Health Service) [2017] QIRC 041

Williams, P. (1996). Suing for negligent teaching: An Australian perspective. *Journal of Law and Education*, *25*(2), 281–306.

11

OUR JOURNEY OF DISCOVERY

Looking back, looking sideways, and looking forward

Introduction

Evidence-informed conversations about graduate quality on program completion are long overdue. In saying this, we recognize that many people express views about the quality of education systems and the effectiveness of teaching practices. Parents and carers want to be confident in the expertise of their child's teacher. They want the teacher to progress their child's learning and prospects for success.

In the revolving doors of teacher education reviews, however, teacher educators have struggled to have a strong, authoritative voice, backed by evidence, showing the quality and effectiveness of programs. Preservice teachers have lacked assurance that there are comparable expectations of quality across the country. In this environment, the professional status of teaching and the attractiveness of the profession have continued to decline.

The 2015 policy move to introduce teaching performance assessments (TPAs) represents what is arguably the most significant reform in teacher education in Australia in recent times (see Chapters 1 and 2). It builds on the introduction of the Australian Professional Standards for Teachers (APST; Australian Institute for Teaching and School Leadership [AITSL], 2011) and the national program accreditation standards (AITSL, 2015). At the time, the idea of TPAs was largely unfamiliar in the teacher education landscape, though there was some knowledge about competence assessment in the United States. Beyond the requirement for demonstrating validity and reliability of the TPA, in Australia, TPAs opened the blue sky: there were no prescribed methodologies for assessing the full cycle of teaching. The challenge was to design and trial authentic assessment of teaching competence across planning, teaching, assessing, and reflecting in a classroom context to determine classroom readiness. The significance of the assessment is perhaps best illustrated by the policy expectation that it would be a valid assessment showing that graduates were

DOI: 10.4324/9780429318504-13

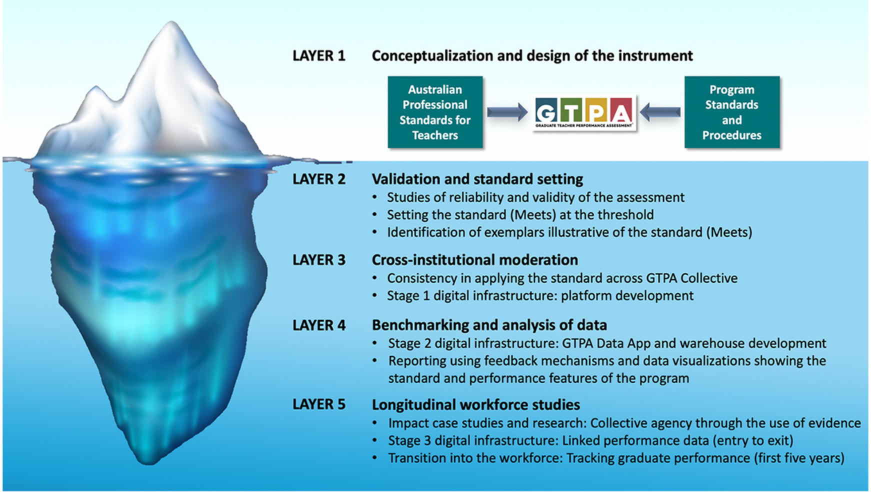
classroom ready on course completion (see Chapter 2 for a discussion on classroom readiness). The term has been actively taken up by some, and actively resisted by others, in part because it was heard as promoting a narrow competence approach to preparation.

So, what has been achieved over the last six years? Twelve TPAs are installed in Australia across the 47 initial teacher education (ITE) providers. Thirty-two of those providers are concentrated in two large consortia.[1] The remaining ten TPAs are spread across 15 providers. However, this development brings with it big questions about the feasibility of establishing a 'ready-to-teach' standard. Much remains to be done.

In the absence of a 'ready-to-teach' standard, there is the real risk of TPA-specific standards of readiness being developed, and by extension, variability in the accepted standard applied across TPAs. Additionally, there are related risks associated with the current absence of coordinated quality assurance systems and processes, both for verifying judgment reliability and more fundamentally, evidence requirements for entry to the profession. In the context of so many TPAs, cross-institutional moderation (CIM) is a further area for attention. This will involve identifying defensible methodologies for monitoring movement in the standard over time and comparability in its application. This observation moves into the space of what has not been achieved over the last six years.

Reflecting how engagement across the country has been patchy, driven largely by the two big consortia, a common vernacular to speak about quality and evidence in TPAs has not yet emerged across the field. Quality assurance processes across TPAs remain largely unarticulated. Within the Graduate Teacher Performance Assessment (GTPA®),[2] however, progress has been made on both fronts. This observation reflects how culture change in teacher education takes time, especially where there is a perception that a top-down policy-driven approach to change has been adopted. When this develops, the act of unifying system validity (the concern of regulatory authorities) and site validity (the concern of teacher education providers including schools) in introducing the TPA reform becomes particularly contested, and misinformation and myths can proliferate. This development also reflects how accountability for teacher education involves state and national agencies, with 'power' for accreditation residing in the former. It is arguably not surprising in these circumstances that the potential of the reform remains unrealized; national conversations about TPAs are long overdue.

Against this background, we discuss the conceptualization of developmental layers that we have chartered to date, and that we identify to be integral to the productive introduction and use of the GTPA (Figure 11.1). Our aim is to convey that the assessment instrument itself is but one layer: easily recognizable 'above the waterline'. However, the potential of TPAs involves collaborative, evidence-informed inquiry into what is 'below the waterline'. Regarding fitness-for-purpose, our starting proposition was that the instrument exists at the intersection of summative assessment (reporting purposes) and formative assessment (improvement purposes); two assessment purposes that have historically operated on competing fronts (see Chapters 4 and 8). These ideas are discussed below.



LAYER 1 **Conceptualization and design of the instrument**

Australian Professional Standards for Teachers
GTPA
GRADUATE TEACHER PERFORMANCE ASSESSMENT
Program Standards and Procedures

LAYER 2 **Validation and standard setting**

- Studies of reliability and validity of the assessment
- Setting the standard (Meets) at the threshold
- Identification of exemplars illustrative of the standard (Meets)

LAYER 3 **Cross-institutional moderation**

- Consistency in applying the standard across GTPA Collective
- Stage 1 digital infrastructure: platform development

LAYER 4 **Benchmarking and analysis of data**

- Stage 2 digital infrastructure: GTPA Data App and warehouse development
- Reporting using feedback mechanisms and data visualizations showing the standard and performance features of the program

LAYER 5 **Longitudinal workforce studies**

- Impact case studies and research: Collective agency through the use of evidence
- Stage 3 digital infrastructure: Linked performance data (entry to exit)
- Transition into the workforce: Tracking graduate performance (first five years)

FIGURE 11.1 Conceptualizing the GTPA as connected layers of research and development

A new conceptualization of teaching performance assessments to realize their potential

In a nutshell, our biggest learning was that a TPA is not a thing, it is a concept. Agreement has to emerge in an ITE community about what it stands for. This stance was informed by a set of key interlocking understandings about preservice teachers, student learning as the core concern of teaching, and the enabling role of assessment.

First is the understanding that preservice teachers are 'already partially constructed' by the teaching they have observed and experienced. In some cases, they are the beneficiaries of these, and in other cases, they may not have thrived. Irrespective of the impact of past classroom experiences, they have nevertheless at least constructed the teacher identity that preservice teachers will adopt in their classroom practice. They have shaped how they present content knowledge, their knowledge of teaching strategies, and their assessment experience. This reflects how they carry with them epistemological beliefs about knowledge formation and knowledge itself, with these tending to remain latent or unarticulated. Similarly, they inevitably bring with them underlying conceptual schemas that have already shaped their attitudes and that will, in turn, shape the choices that they make as teachers and those they offer to students in learning.

The challenge for a TPA therefore is to focus preservice teacher attention on the complete triad of teaching: *what* (e.g., lesson planning using content knowledge); *how* (e.g., choices of teaching strategies); and *why* (e.g., the reasoning and decision-making before, during, and after teaching that shape practice and interactions with students). The focus on *why* leads to scrutiny of the constituent processes and activities that underlie and motivate actions, talk, and surface behavior (Phelps, 1989,

p. 37) occurring in classrooms. Unless these are problematized through an inquiry process, they can readily become naturalized and taken-for-granted as how teaching should occur in a particular school.

Second, it is time to bring student learning to the center of both the academic and practical programs of teacher preparation. Our advocacy for this position reflects the understanding that student learning has tended to remain in the shadows of teaching and assessment. This is reflected in the emphasis preservice teachers give to questions such as: What am I going to teach? How am I going to teach it? At least of equal importance are questions such as: How will I know what students have learned? How will I discern the barriers to student learning? How will I address these through teaching and formative assessment feedback? With this turn, the aim is for preservice teachers to come to understand that previously taken-for-granted understandings about knowledge and knowledge formation "become problematic and subject to inquiry" (Phelps, 1989, p. 44).

In the GTPA Collective, we also decided to work on two fronts that permitted reach into (1) summative reporting of classroom readiness (for preservice teachers) and program effectiveness (for universities); and (2) formative use of data generated through CIM for curriculum review and program renewal. While the two purposes of assessment have been historically distinguished and separated, in our approach, we saw merit in centering on purpose and exploring, in a large group, how data collected could be used for reporting and for improvement purposes. Figure 11.1 shows the design of the GTPA instrument as above the waterline, with four other layers (aspects of development) appearing below the waterline. While the latter may be less visible, they are nevertheless essential for achieving both purposes and building the infrastructure to promote and sustain quality programs in ITE. The lessons below attempt to present what we have learned with and through the GTPA Collective.

Lesson 1: Conceptualization and design of the instrument (Layer 1)

Professional competence assessments, including TPAs, are not *just another assessment*. TPAs are located in the nexus of theory and practice. Generally speaking, they involve demonstrating practice and articulating reasoning and instructional decision-making in action, including connecting theory, research, and policy. This requires that a TPA is recognized and accepted by the profession.

Preceding the introduction of TPAs, teacher educators had adopted portfolios to demonstrate professional growth and development toward *being* a teacher. Typically, portfolios were a collection and collation of evidence over time, with commentary on events and interactions. Evidence and commentary were intended to demonstrate professional standards. Portfolios could be considered an approach to *performance assessment over time*.

A TPA in Australia is a demonstration of practice in situ and is to be completed in the final year of preparation during a sustained placement in a classroom context. Through successful completion of this assessment, preservice teachers and the public are to have confidence that teaching graduates are prepared to make the transition into the profession, that is, those entering the profession have demonstrated core skills recognized by the profession. This reflects one of the original aims of the assessment, namely, to take a crystalline focus on the transition from teacher preparation in universities to entry into the workforce.

The GTPA is a *performance assessment at a point-in-time* – program completion. It requires preservice teachers to not only demonstrate practice, but also address the efficacy of their practice – its impact on learners. Attention shifts from teaching to learning to address questions such as: What learning has occurred? How will I know what learning has occurred? How do I use formative assessment evidence to progress next-step teaching?

Teacher educators have reported that a focus on the use of evidence to inform preservice teacher classroom practice has strengthened their decision-making capabilities (see Chapter 9). They have also reported their confidence in the assessment acting as a valid and reliable tool to support their responsibilities as gatekeepers to the profession. The design of the instrument and the accompanying scoring rubric have provided a vernacular that both teacher educators and preservice teachers use to speak about practice and address issues of quality, evidence, and data to improve teaching for whole class groups and individual students.

Lesson 2: Validation and standard setting (Layer 2)

The critical function of the TPA as a proxy for profession readiness calls for rigorous processes for validation of the assessment and for establishing the standard at Meets (see Chapters 4 and 6). This includes a validated scoring rubric that accompanies the TPA, with evidence of demonstrated reliability and exemplars (authentic TPA samples) showing the range of quality. The exemplars are concrete illustrations of what the standard looks like in practice. There must be a sharp focus on samples illustrating Meets (at the threshold) and Does Not Meet (just below the threshold).

Once there is a validated assessment and a defined standard of Meets with accompanying exemplars, clearly stated and defined conditions need to be in place for addressing ongoing fidelity in TPA implementation (see Chapter 7). Over the past six years, discussions of standards, expected features of quality (criteria), and processes for scoring and arriving at overall judgment have been ongoing. This was expected, noting the tight connections to be demonstrated between the TPA and professional and program standards, as discussed elsewhere (see Chapter 3).

While teacher educators bring with them their evaluative experience and expertise, it would be naïve to expect that there was a common standard in teacher education prior to the introduction of the TPA. Validation of the instrument, implemented at scale, and rigorous standard-setting processes, results in initial evidence of 'the line in the sand' – what meeting the standard looks like. Longitudinal monitoring of the standard can be done both within and across universities using anchor samples. Unless attention is paid to capturing this movement, the case could be made that it was easier to graduate from teacher education last year or the year before, or from a different university. This brings into play issues of fairness in tests of graduate readiness to enter teaching (see Chapter 5).

This monitoring has the potential to build a longitudinal evidence base to show the quality and effectiveness of teacher education at regional, state, and national levels. It would be naïve to assume that the implementation of a TPA, in and of itself, leads to judgment reliability. In our experience, the latter is dependent on rigorous quality assurance systems and processes, including CIM.

Lesson 3: Cross-institutional moderation online (Layer 3)

Teacher educator participation in CIM is foundational to the goals of assuring graduate teachers are well-prepared for the classroom (see Chapter 7) and for building a longitudinal evidence base, as mentioned. A weakly framed position on CIM and evidence undermines the prospect for delivering the promise of the Teacher Education Ministerial Advisory Group review (TEMAG; Craven et al., 2014) to improve the quality of graduate teachers and professionalize teacher education.

A key outcome from the GTPA project to date includes new evidence showing that ITE programs are of varying quality across the country, and that internal moderation alone, as undertaken within individual universities, does not necessarily deliver comparability in the application of an established standard (see Figure 11.4 and related discussion). Standards typically exist in the head, that is, in latent or unarticulated form (Sadler, 1987). It is potentially difficult for an assessor to unearth and articulate the bases of judgment (Adie, 2014; Phelps, 1989; Smith, 1995; Wyatt-Smith & Klenowski, 2013). The move to CIM-Online™[3] in the GTPA project was necessary to build knowledge of the standard established and used in universities within the GTPA Collective located across jurisdictions.

The data literacy of teacher educators is therefore core business. The purposeful inclusion of practices, processes, and resources to develop and sustain a shared understanding of expected characteristics of quality is necessary to support professional judgment. Calibration training, customized decision aids, illustrative exemplars, and descriptive commentaries on how an established standard is applied are essential in building judgment dependability.

Further, over time, CIM-Online™ and calibration training have the potential to build confidence among stakeholders that an agreed level of quality is being applied to all candidates entering the workforce, irrespective of the university in which preparation was undertaken. Without a focus on comparability and rigorous mechanisms for demonstrating that a comparable standard is applied across teacher education providers, the nation could simply revert to each university applying a university-specific standard. This would be an undesirable step back from the prospect of establishing agreed expectations of quality across the country; the standard on completion of a program should speak to profession readiness. Equally undesirable would be a move to standardize programs or to install a single TPA.

Going to scale calls for investment in digital infrastructure

Generating evidence from data collected across universities, separated by vast geographical distances, requires digital infrastructure. It is not enough to ask teacher educators to submit data without support for how to do this. Enabling infrastructure in the form of online platforms for submitting TPA samples, web-portals for accessing and scoring samples in CIM, software apps for cohort data submission, and the digital generation of reports are a necessary provision. This requires specialist development work by a multidisciplinary research team with complementary skills.

The EQuITE data warehouse (see Chapter 7) has been structured to enable the repeated annual intake of online scoring data, as well as individual preservice teacher GTPA performance outcomes and contextual information for entire ITE program cohorts across universities in the Collective. This includes data across a range of demographic characteristics such as program type, discipline area, mode of delivery, phase of schooling, and placement variables. The data provide evidence of where problems exist and can inform the solutions. This enables future analyses to examine longitudinal trends applying the GTPA established standard within and across universities, as discussed below.

Figure 11.2 depicts the three main phases of CIM implementation that have been developed to support the GTPA. The phases, as outlined, have relevance whether CIM is conducted with large, geographically dispersed teams or with smaller, in-person moderation meetings. The CIM activities have been elaborated in Chapters 7 and 8 in this volume. Readers may be interested in related discussions in Wyatt-Smith et al. (2021) that present accounts of teacher educators and researchers involved in CIM.

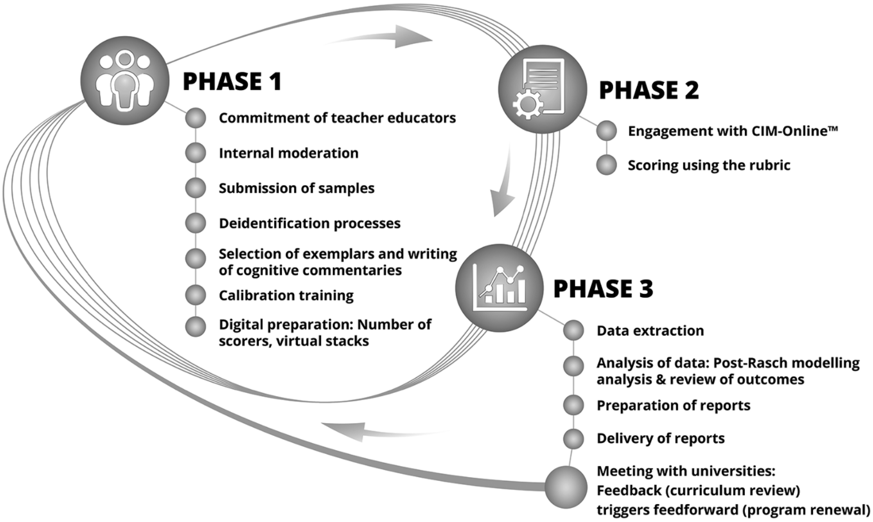

PHASE 1
Commitment of teacher educators
Internal moderation
Submission of samples
Deidentification processes
Selection of exemplars and writing of cognitive commentaries
Calibration training
Digital preparation: Number of scorers, virtual stacks
PHASE 2
Engagement with CIM-Online™
Scoring using the rubric
PHASE 3
Data extraction
Analysis of data: Post-Rasch modelling analysis & review of outcomes
Preparation of reports
Delivery of reports
Meeting with universities: Feedback (curriculum review) triggers feedforward (program renewal)

FIGURE 11.2 Annual phases of benchmarking and reporting in the GTPA

Lesson 4: Benchmarking and analysis of data (Layer 4)

At this critical juncture in the history of teacher education in Australia, we must avoid the prospect of a tiered system of teacher preparation. Put simply, we have to get TPAs right. It is time for Australia to make transparent a common or agreed standard of graduate readiness that applies irrespective of location, mode of delivery, or ITE program. Incentivizing groups of universities to work together in CIM to demonstrate comparability in scoring is a necessary precondition for moving to a more ambitious enterprise of benchmarking teacher education nationally.

In our work to date, we have applied a best-practice methodology for benchmarking the quality of graduate performance. In this venture, data visualization and reporting of customized results to participating universities are presented to show:

1. How the established standard has been applied in each university program.
2. How the stated criteria (scoring rubric) have been applied at the cohort level in each program.

Previously, standards have been used much like checklists to structure and review program design. As such, they have tended to act as inputs into teacher education for program planning. They allow easy tracing to see where knowledge or skill was taught, practiced, and assessed. Through the analysis of actual TPA scoring, evidence of standards 'met' can be used summatively, to show preparedness for workforce entry, and formatively, for program renewal. This forms a feedback loop that connects standards and evidence to quality assure graduate readiness, as well as inform reflection on the quality and impact of ITE programs (see Chapter 8).

Among participants of the GTPA Collective, evidence of demonstrated comparability in what counts as the passing standard in ITE is a non-negotiable expectation, tightly held. This perspective is directly evident in how they talk about confidence in the assessment and fairness in how it is assessed (see Chapter 9). In long-term exchanges with members of the Collective, we have also observed first-hand that data only has value if teacher educators actively interpret it, infer meaning from it, and use it in context. For this to occur, their input into decisions about data visualization approaches is essential.

Modeling interpretations of data by experts is helpful in building teacher educators' data literacy. For most, the type of data coming from the GTPA is not of a type that was part of their own preparation to be a teacher, or their doctoral research programs. To address these observations, the Collective served as a community of learners where teacher educators shared interpretations of data and how they could apply these in decision-making. More than this, the acts of sharing data broadened the professional learning circle to include members of regulatory authorities, policy personnel, school leaders, teachers, and researchers. The discerning use of data as core business has been taken up as a shared enterprise, becoming more than a terminal report card. It has stimulated cross-university projects as well as intra-university research to inform learning and teaching.

Lesson 5: Longitudinal workforce studies (Layer 5)

The deepest level of inquiry involves teacher education research at scale. Building on GTPA performance data and data from the practical program undertaken in schools, we have initiated longitudinal studies examining candidates' performance trajectories over the duration of preparation. In short, we follow candidates from entry, throughout progression in the degree programs, at Bachelor and Masters levels, and into the workforce. Using data on individual candidates, we examine performance data, including timing, attempts, and outcomes, to shed light on the barriers facing cohorts of special interest and the points at which risks of separation from programs emerge and become acute. The potential of this research includes new knowledge about performance progression and the sufficiency of entry standards. It also provides opportunity to generate new knowledge about the optimum timing of customized supports to enable successful completion and, in so doing, address attrition. Driving this, in part, is recognition of the acute workforce shortages in teaching already felt in Australia and reported internationally.

In designing the longitudinal data-linking study, we adopted the position that no single assessment method can provide all the data required for inquiring into progression. We need to see the suite of assessments, timepoints, and scores to see the development of the teaching graduate over time. In teacher education, we currently have boxes of evidence that can tell us something about our teacher candidates, but they are typically discrete and have not been linked to tell the stories about cohorts and individual pathways.

Further, while there are professional standards that underpin ITE programs, scant attention has been paid overall to important aspects of preparation such as professional dispositions and how these are shaped during candidature. How aspects of practice and dispositions are fostered within teacher education programs and how they are brought together is an important next piece for policy, practice, and research.

Closing commentary: Looking back, looking sideways, looking forward

Looking back

For this project and the GTPA Collective, we started the journey with the centerpiece being student learning and quality teachers to deliver the best possible outcomes for their students. Our experiences have been captured as a theorization of practice and shared through the practical application of change through collaboration.

As we began this journey, it was clear to us that following the release of the TEMAG report (Craven et al., 2014), change was to occur in teacher education and importantly, it was to be externally imposed, driven by policy, and bringing with it a potentially narrow accountability focus. Taking up the vantage point of research, we saw the introduction of a TPA as an opportunity to inquire into the quality and impact of teacher education programs. We recognized the widely reported finding that teacher education lacked an evidence base to present claims of quality and effectiveness. It was therefore ill-equipped to engage productively with growing public concerns with the quality of graduates which in turn brought negative impacts on the status of the profession. The recruitment of high-performing candidates was becoming increasingly difficult. It was therefore an imperative to take up the challenge to offer a coherent theorized and empirically validated response to the call for introducing a TPA. In retrospect, our decision was to take the unchartered road and begin, what we have referred to earlier, as our journey of discovery. Along this journey, we have:

1. Mobilized large-scale collaboration and partnerships that challenged entrenched geographic and disciplinary silos in teacher education preparation. We continue to collaborate across a large and growing group and scaling up has made all the difference. This has extended to our digital infrastructure and data processes.
2. Applied judgment and decision-making methodologies and data analytics that had not previously been applied in teacher education. We continue to work from the position that to fix a problem, we need to see it and fully understand it. Evidence is therefore critical.
3. Designed and implemented a customized approach to cross-institutional moderation online (CIM-Online^{TM}) for benchmarking across multiple teacher education institutions. We continue to build teacher educators' evaluative expertise and data literacy, working across states and territories.
4. Designed and implemented longitudinal investigations into graduate readiness and program effectiveness, taking a dual focus on candidature progression through to entry into the teaching workforce. We continue to work with universities and industry partners, including government agencies at state and national levels, to inform teacher education policy.

Looking sideways

Here, we present two examples of change stemming from the GTPA and the work of the Collective; both show evidence of culture change in teacher education. The first example, discussed below, relates to teacher educators' judgment calibration training and subsequent participation in moderation as taken up by a large group of teacher educators with no prior history of working together. Preconditions for this to occur included the build of trust in the Collective – trust in the benefits of a participatory model of research. This was made possible because the community of teacher educators held a shared motivation to improve teacher education through their actions and their desire to take up agency in building an evidence base to show the effectiveness of their programs. Another necessary precondition was the support of university leaders in faculties of education to provide human and material resources necessary for building the evidence base. Teacher educator evaluative expertise was central in this shared enterprise.

The second example, discussed below, relates to benchmarking (see Layer 4 above and Chapter 9). This was made possible through teacher educators' commitment to online scoring. This reflects how moderation and calibration became a subject of systemic inquiry into scoring and the application of the GTPA standard to be applied across the Collective. Judgment dependability was a shared proposition for how the group worked; shared trust in CIM-Online™ was integral to quality assurance systems and processes. Motivating the group was the opportunity to see evidence of the quality of their graduates in new ways, locally within jurisdictions and across the nation. The problem of calibration, moderation, and how the two relate to build dependability of teacher judgment came to the center as a systematic inquiry over the period of 2018–2021. This opened the space to investigate judgment comparability.

EXAMPLE 11.1 INCREASING ENGAGEMENT IN CALIBRATION AND MODERATION TO BUILD JUDGMENT DEPENDABILITY

As shown in Figure 11.3, data from 2018–2021 CIM-Online™ reveals a significant uptake in teacher educators' participation in calibration and moderation events. These outcomes show that the number of judges involved in calibration or moderation events increased approximately three-fold between 2018 and 2021. There has also been an increase in the number of judges who participated in both calibration and moderation. This pattern of growth points to an increase in teacher educators' understanding of the importance of calibration training in building judgment dependability across the Collective.

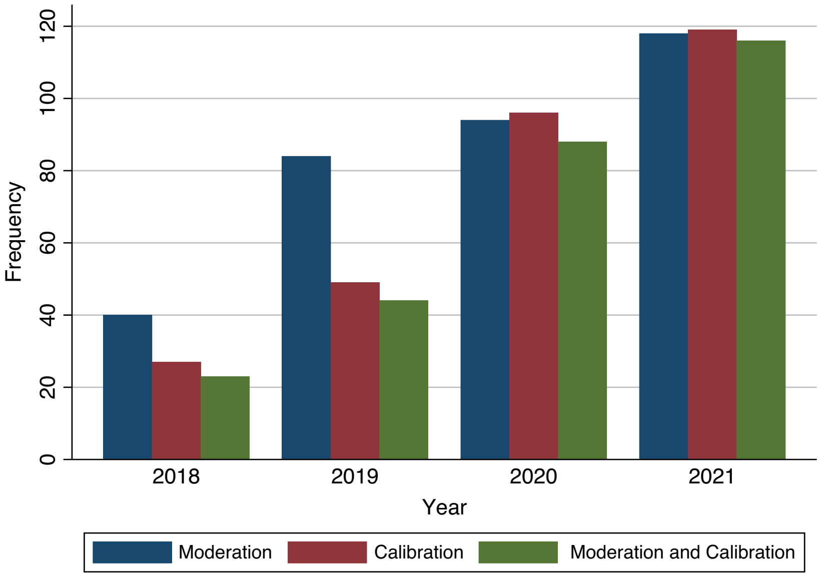


FIGURE 11.3 Number of judges who participated in moderation and calibration between 2018 and 2021

Analyses from 2018–2021 CIM-Online™ also revealed an association between increasing participation in calibration *and* endorsement[4] of pre- and post-moderated scores.

EXAMPLE 11.2 IMPACT OF CALIBRATION AND MODERATION ON ENDORSEMENT

Further analysis was undertaken to investigate the change in endorsement over time. Figure 11.4 visualizes predicted probabilities of endorsed samples in each university across the period 2019–2020. These results are derived from a multilevel logistic regression model for sample endorsement, where samples are grouped by university. Figure 11.4 is presented in two parts for ease of viewing, grouping universities together based on the predicted probabilities in 2019 (above and below 0.8). Each line represents the change in endorsement over the period 2019–2020 for a particular university, with the dashed line representing the change in endorsement over time across all universities.

There are three findings: (1) The average endorsement increases significantly over time, as presented by the dashed line in Figure 11.4; (2) the change in endorsement over time is not uniform: while different levels of improvement in endorsement are observed for most universities, there is one exception (University B); (3) the variation of sample endorsement across universities is considerably smaller in 2020 than in 2019. As mentioned, this positive change points to the benefits of calibration training and a sustained focus on building judgment reliability.

To revisit the two core questions – What has changed? What has been achieved? – we have generated new evidence to make quality visible. Carrying this forward, we have engaged a national professional community to inquire into the quality and effectiveness of teacher education using this evidence. We have shone the spotlight on the nature and function of a standard to capture graduate quality at the Meets level and also the contribution of criteria to see cohort characteristics of performance. Working with the Collective, we have brought forward evidence of how the standard and criteria apply, both for interrogation and subsequent use in a diverse range of programs and contexts. We have concentrated on analyzing evidence and inferring meaning from and using the results of the analyses. This has involved co-constructing with teacher educators new forms of data visualization in teacher education. It has also involved providing feedback on proposed forms of representing the data and discussion of their potential utility to inform curriculum review and program renewal. Such endeavors have promoted data literacy with direct relevance to pedagogy in teacher education. As teacher educators expand their repertoire of practices using data, they have opened new ways to 'see' quality and in turn inquire into claims about quality. This information has not been previously available to teacher educators, nor did they have access to ways to build such an evidence base.

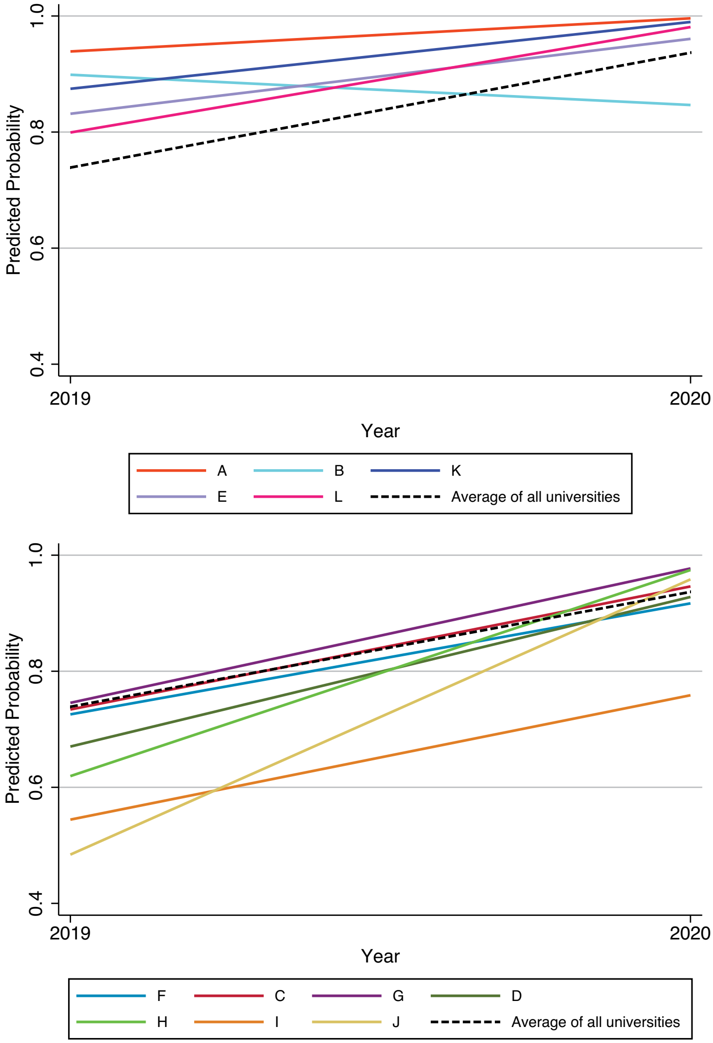


FIGURE 11.4 Predicted probabilities for endorsement by universities over the period 2019–2020. For visualization purposes, universities were grouped based on their predicted probabilities in 2019: Above and below 0.8

Complementing the focus on the inner workings of teacher education programs, we have begun an inquiry into the inner workings of teacher educators' judgment and decision-making. This too represents previously unchartered territory. This has been achieved by focusing on data at two levels: (1) The level of the standard to look at quality performance (Meets, Above, and Below) and (2) using criteria to look at characteristics of performance in cohorts. This data can be used in strengthening the teacher educator community in how to discern quality in ITE.

Looking forward

We identify four essential next steps for realizing the potential of TPAs in effective and sustainable approaches to promote ITE reform. First, recognition is to be given to TPAs as a valued tool for reforming ITE. There is general agreement in policy that we need to get TPAs right; however, the nature and function of TPAs themselves are not well known or understood among the teaching profession and the wider community. Currently, the evidence that TPAs produce is not widely valued. While the uptake of a TPA has been an additional requirement on teacher education, there has been no additional funding. Similarly, there has been no recognition of the significant human and material costs borne by some universities.

Second, a *prepared-to-teach standard* that applies to TPAs should be defined. In a context where there are a significant number of TPAs, it is likely that there will be differences among the overall pass standard. Referring to the GTPA and as discussed in Chapters 6 and 7, quality assurance systems and processes have been developed to support the implementation of the passing standard for this assessment. However, discussion of how performance expectations or the standard in this assessment relate to those in other TPAs is yet to begin.

Third, the various TPAs in the country are at different stages of maturation. The maturing of TPAs is an essential pre-condition for realizing their potential, both for summative purposes and formative purposes, including curriculum review and program renewal. In this process, it has become clear to us that the focus on workforce transition is a critical next step to examining the impact and effectiveness of the preparation program.

Fourth, content area TPAs in areas of national priority, and phase appropriate TPAs, including the early years should be investigated. This would support a strengthened focus on the teaching of curriculum, addressing the literacy and numeracy demands to support learning in curriculum areas. We have much to learn about how to capture and analyze authentic performance data in applied disciplines and efficient ways of deidentifying such data, especially as it involves video footage of students. This is foundational to tracking the movement of the standard over time, in part through using anchor samples. Legal precedence for cases contesting grading decisions in the case of the Educative Teacher Performance Assessment (edTPA) and Performance Assessment for California Teachers (PACT) is instructive for Australia (see Chapter 10).

Beyond these four steps, there are thorny issues that merit attention. While not offering a comprehensive list, we suggest that these include set endorsement periods for TPA implementation, and the provision of performance data from a TPA at defined intervals. This would serve to show not only how TPAs are implemented in the field, but also to gauge their impact in leveraging improvement in ITE.

These observations point to a necessary shift from having a TPA to the evidence that a TPA produces. The spotlight then moves to uses of evidence including, for example, sustaining a culture of inquiry and improvement in ITE. A well-developed system of quality assurance processes is part of this move. Exploration of the nature and function of CIM (see Chapter 7) in education policy also merits attention.

Currently, the key role of moderation in maturing TPAs and in furthering the professionalization of the teaching workforce is in its infancy, as is the role of the teacher educator in program evaluation. For optimal effect, this role would engage teacher educators and colleagues in the schooling sector, including principals and mentor teachers.

In conclusion, there can be no doubt that our journey of discovery involved risk-taking on the part of all involved in the GTPA Collective and multidisciplinary research team. For the first time, we were subjecting our assumptions and thinking about quality and methodologies, to examination in a large scholarly community. Together, we have begun building an evidence base for teacher education using digital infrastructure and systems thinking. While this work continues, we are extending into workforce studies building a longitudinal evidence base showing candidature pathways through ITE and into teaching.

We have come to understand that the TPA is not a thing, it is a concept – agreement has to emerge in an ITE community about what it stands for. Over the period of this journey, it is our experience that shared conceptions through collaborative inquiry evolve into deeper and more comprehensive understandings of quality in teacher education. These have fueled our understandings of the role of TPAs and their potential for improving teacher education.

Notes

1 The two large consortia are the Assessment for Graduate Teaching (A*f*GT) based at The University of Melbourne (see https://education.unimelb.edu.au/research/projects/assessment-for-graduate-teaching-afgt) and the Graduate Teacher Performance Assessment (GTPA) based at the Australian Catholic University (see https://www.graduatetpa.com/).

2 Acknowledgment: The Graduate Teacher Performance Assessment (GTPA®) was created by the Institute for Learning Sciences and Teacher Education (ILSTE), Australian Catholic University (ACU), and has been implemented in a consortium of Australian universities, known as the Collective (graduatetpa.com).

3 Acknowledgment: The online model of cross-institutional moderation (CIM-Online™) was conceptualized and developed in the Institute for Learning Sciences and Teacher Education, Australian Catholic University. For a discussion of CIM-Online™, readers are advised to also see Wyatt-Smith et al. (2021).

4 Endorsement refers to the agreement between internally moderated scores as submitted by universities (pre-moderated) and cross-institutionally moderated scores as determined by the GTPA Collective in CIM-Online™ (post-moderated).

References

Adie, L. E. (2014). The development of shared understandings of assessment policy: Travelling between global and local contexts. *Journal of Education Policy*, *29*(4), 532–545. https://doi.org/10.1080/02680939.2013.853101

Australian Institute for Teaching and School Leadership (AITSL). (2011; revised 2018). *Australian professional standards for teachers.* https://www.aitsl.edu.au/teach/standards

Australian Institute for Teaching and School Leadership (AITSL). (2015; revised 2018, 2019). *Accreditation of initial teacher education programs in Australia: Standards and procedures*. https://www.aitsl.edu.au/docs/default-source/national-policy-framework/accreditation-of-initial-teacher-education-programs-in-australia.pdf?sfvrsn=e87cff3c_28

Craven, G., Beswick, K., Fleming, J., Fletcher, T., Green, M., Jensen, B., Leinonen, E., & Rickards, F. (2014). *Action now: Classroom ready teachers*. https://docs.education.gov.au/documents/action-now-classroom-ready-teachers-report-0

Phelps, L. W. (1989). Images of student writing: The deep structure of teacher response. In C. Anson (Ed.), *Writing and response: Theory, practice, research* (pp. 37–71). National Council of Teachers of English.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*(2), 191–209. https://doi.org/10.1080/0305498870130207

Smith, C. (1995). Teachers' reading practices in the secondary school writing classroom: A reappraisal of the nature and function of pre-specified assessment criteria. Unpublished doctoral thesis. University of Queensland.

Wyatt-Smith, C., Adie, L., & Nuttall, J. (Eds.) (2021). *Teaching performance assessments as a cultural disruptor in initial teacher education: Standards, evidence and collaboration*. Springer.

Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: Types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 35–52. https://doi.org/10.1080/0969594X.2012.725030

GLOSSARY

In producing this Glossary, we have drawn on the writing of the authors of the book, along with other publicly available definitions and descriptions. The latter include teacher education agencies, government policy materials concerning teacher education, and other authors (see references below).

ACCREDITATION Initial Teacher Education (ITE) in Australia must meet the program standards developed by the Australian Institute for Teaching and School Leadership (AITSL). These are the "Standards and Procedures that set out the requirements that an initial teacher education program must meet to be nationally accredited… They are designed to ensure that all graduates of initial teacher education meet the Australian Professional Standards for Teachers at the Graduate career stage" (AITSL, 2015, p. 3). (See also *Initial Teacher Education [ITE]* and *Program Standards* for related terms).

ANCHOR SAMPLE A sample selected to illustrate a designated level of quality.

AUSTRALIAN PROFESSIONAL STANDARDS FOR TEACHERS (APST) The Australian Professional Standards for Teachers (APST) were developed by the Australian Institute for Teaching and School Leadership (AITSL) and comprise of seven Standards. These standards establish "what teachers should know and be able to do" (AITSL, 2011, p. 4). They are "interconnected, interdependent and overlapping" (AITSL, 2011, p. 4). The Standards consist of three domains of teaching: (1) professional knowledge, (2) professional practice, and (3) professional engagement. (see also *Professional Standards* for related terms.)

BENCHMARK A level of quality against which performance can be measured.

BENCHMARKING The practice of comparing processes and performance metrics, often for quality assurance purposes.

CALIBRATION Calibration includes decision aids that support the judgment process. These include (1) exemplars selected to illustrate characteristics of the standard and (2) cognitive commentaries that provide the basis of judgment, including how types of performance were combined in arriving at an overall judgment.

COGNITIVE COMMENTARIES Cognitive commentaries are written specifically for each sample using the criteria specifications and the performance level descriptors (PLDs) as a basis for the descriptions of performance. That is, the cognitive commentary is written using the specific features of the performance rather than the general level of description in the criteria or descriptors. Where necessary, raters receive a second set of samples for further calibration. In this case, the cognitive commentary is provided with the sample, so that raters are trained to see critical evidence identifying a level of performance (see also *performance level descriptors [PLDs]* for related terms).

CRITERIA/ CRITERION "A standard of judgment or criticism; a rule or principle for evaluating or testing something" (dictionary.com).

CROSS-INSTITUTIONAL MODERATION (CIM-ONLINE™) CIM-Online™ involves the recording and collation of judgment decisions with the use of an established standard and relies on digital technologies and online scoring systems. It involves a blind review process that requires assessors to rate and record scores on authentic samples provided by multiple ITE institutions. The samples are fully de-identified, including the deletion of the original score provided by the host institution.

DATA WAREHOUSE An online repository for storage or archiving and is a system for housing data from a variety of sources. In the GTPA, this is a purpose-built digital data system for storing de-identified performance records. The records are collected to (1) monitor the movement of the standard over time; (2) enable longitudinal investigations into the quality of ITE programs and their effectiveness, and (3) study the characteristics and performance trajectories of individuals and sub-cohorts of special interest.

DEPENDABILITY Dependability of judgments has been located at "the intersection of reliability and content validity" (Wiliam, 1994, p. 18). There are five actions used to investigate dependability: "the specification of the tasks; the specification of the criteria; training; moderation; and the development of an 'assessment community' within the school allied to increased confidence in the professional judgment of teachers" (Harlen, 2004, p. 28). Concerns about dependability have been associated with the influence of bias and varying interpretations of the meaning of criteria and standards (Harlen, 2004).

edTPA (EDUCATIVE TEACHER PERFORMANCE ASSESSMENT) The "edTPA is a performance-based, subject-specific assessment and support system used by teacher preparation programs throughout the United States to emphasize, measure and support the skills and knowledge that all teachers need from Day 1 in the classroom" (edTPA, n.d, n.p).

ENDORSED SAMPLE A sample is considered to be endorsed when the outcomes of scoring and analysis confirm the score recorded on the sample at the point of initial submission.

ENDORSEMENT The agreement between internally moderated scores as submitted by universities (pre-moderated) and cross-institutionally moderated scores as determined by the GTPA Collective in CIM-Online™ (post-moderated).

EVIDENCE FOR QUALITY IN INITIAL TEACHER EDUCATION (EQUITE) Collectively, this system, known as the Evidence for Quality in Initial Teacher Education (EQuITE), forms a digital architecture designed for the core purpose of providing feedback for curriculum review and program renewal. The data generated through EQuITE are used to build an evidence base on the effectiveness of teacher education, and to improve the quality of teacher education through networks and collaborative partnerships across universities, sector authorities, and levels of government at state and national levels in Australia.

EXPERT MODERATION Where the evaluative experience of a nominated expert is used to make the final judgment decision.

FIDELITY The extent to which an assessment is implemented as intended. Fidelity is pivotal in efforts to safeguard fairness for preservice teachers so that all preservice teachers learn and provide evidence of their knowledge, skills, and decision-making. (For related concepts, see system validity and site validity.)

GRADUATE TEACHER PERFORMANCE ASSESSMENT (GTPA®) An officially endorsed Australian teaching performance assessment conceptualized and designed by the Institute for Learning Sciences and Teacher Education, Australian Catholic University. It is implemented in partnership with a national Collective of Australian universities (https://www.graduatetpa.com/).

GTPA COLLECTIVE Is the consortium of teacher educators from across Australian universities that have chosen to apply and use data produced by the GTPA in programs at both Bachelor and Master degree levels (https://www.graduatetpa.com/discover/).

INITIAL TEACHER EDUCATION (ITE) ITE is a degree program (Bachelor and Master levels) and is offered at universities or accredited higher education colleges. It is a tertiary-level course/program undertaken by teacher education candidates (also known as preservice teachers) to build the knowledge and skills required for qualification as a registered teacher.

INTELLIGENT ACCOUNTABILITY Acknowledges the professional responsibility teachers have for student learning and their personal responsibility for self and community within the profession.

LITERACY AND NUMERACY TEST FOR INITIAL TEACHER EDUCATION (LANTITE) A compulsory assessment to be completed by all ITE graduates prior to graduation to assess their personal literacy and numeracy skills. The test contributes to the promotion of public confidence in graduates of teacher education programs (https://teacheredtest.acer.edu.au/).

MODERATION A practice that demonstrates reliability and comparability of scoring and related judgments and contributes to quality assurance systems and processes. There are a variety of forms of moderation that include both statistical moderation and social consensus moderation. Statistical moderation is more commonly practiced in examination systems. Social consensus moderation involves teachers/raters typically working in small teams to discuss and review their judgments by the use of an established benchmark or common standard.

NATIONAL PROGRAM STANDARDS (AITSL) The requirements set by the Australian Institute for Teaching and School Leadership (AITSL) for the development of Australian ITE programs. These are for national accreditation purposes with accreditation the responsibility of state regulatory authorities (https://www.aitsl.edu.au/deliver-ite-programs/standards-and-procedures).

ONLINE MODERATION Online moderation involves synchronous and asynchronous processes. In the context of the GTPA, teacher educators use a common rubric and an established standard to score performances and record judgment decisions online (see *Cross-institutional moderation* [CIM-Online™] as a related concept).

PAIRWISE COMPARISON Comparison of two work samples to identify which is better.

PERFORMANCE ASSESSMENT A performance assessment, also referred to as a *complex performance assessment*, is the demonstration of the decision-making, skills, practices, and knowledge as recognized in the performance of professionals in a given profession or field.

PERFORMANCE ASSESSMENT FOR CALIFORNIA TEACHERS (PACT) "The PACT was the teaching performance assessment used in 32 IHEs and internship programs that were customized to seventeen different credential areas. By completing PACT, teaching candidates created a Teaching Event that was an extended documentation of a segment of student teaching. Integrated across the domains of teaching, Planning, Instruction, Assessment, Reflection and Academic Language, PACT required candidates to demonstrate both content pedagogical knowledge and higher order thinking skills." (https://scale.stanford.edu/teaching/pact)

PERFORMANCE LEVEL DESCRIPTORS (PLDS) The descriptors identify qualities of performances on the GTPA expected of a graduate at Meets, Above, and Below.

PORTFOLIO A collection and collation of evidence over time with commentary on events and interactions, often associated with the required professional standards.

PRESERVICE TEACHER (PST) A candidate in an initial teacher education course. Preservice teachers must complete both academic and practical (field) requirements and achieve a minimum of a passing grade on both requirements for graduation and licensure. In the Australian context, preservice teachers work toward a degree qualification (Bachelor or Master).

PROFESSIONAL EXPERIENCE The practical component of a teacher preparation course and undertaken in a school. It is also referred to as *practicum* and it is distinct from the academic component. Professional experience is often scheduled throughout an ITE course and is often completed over several weeks.

PROFESSIONAL STANDARDS Established and developed by the Australian Institute for Teaching and School Leadership (AITSL), the Australian Professional Standards for Teachers (APST) comprise of seven Standards. These standards establish "what teachers should know and be able to do" (AITSL, 2011, p. 4). (See also *Australian Professional Standards for Teachers [APST]* for related terms.)

PROGRAM STANDARDS Established and developed by the Australian Institute for Teaching and School Leadership (AITSL), these are the "Standards and Procedures that set out the requirements that an initial teacher education program must meet to be nationally accredited…They are designed to ensure that all graduates of initial teacher education meet the Australian Professional Standards for Teachers at the Graduate career stage" (AITSL, 2015, p. 3). (See also *Accreditation* for related terms.)

RATER(S) A teacher educator who assesses the GTPA.

RELIABILITY An agreement in rater scores at the criterion level, and an adjusted measure of inter-rater agreement.

RUBRIC The GTPA scoring rubric was designed to align with the core practices of *planning, teaching, assessing, reflecting*, and *appraising* in response to the Graduate Teacher Standards (AITSL, 2011) and the Program Standards (AITSL, 2015). It is an established mode of protocol for assessing the GTPA.

SAMPLE An original copy of a completed GTPA submitted by a preservice teacher considered by the university to represent a level of quality.

SITE VALIDITY Recognizes that the assessment instrument and related performance criteria (rubric) are fit-for-purpose as measured against *local site requirements*. This can include the practices intended that are responsive to local or community contexts of a school where a preservice teacher has completed their practicum or as it pertains to a teacher education program. Site validity recognizes the local influences of practices valued in a specific school or community setting.

STANDARDS In Australia, there are two distinct standards for teachers and ITE to meet. These are the *Professional Standards* and the *Program Standards*. Both were developed by the Australian Institute for Teaching and School Leadership (AITSL) (see also *Australian Professional Standards for Teachers* [APST] for related terms).

SUPERVISING TEACHER This refers to a fully qualified practicing teacher in a school who supervises a preservice teacher's professional experience in a school setting.

SYSTEM VALIDITY Recognizes that an assessment instrument and related performance criteria (rubric) are fit-for-purpose as measured against *official/system requirements*. The APST should be evident in both the design of the assessment and in the generation of evidence of professional competence.

TEACHER EDUCATION MINISTERIAL ADVISORY GROUP (TEMAG) Formed in 2014 with the role of advising the Australian Federal Government on how initial teacher education courses can guarantee that graduating teachers have the most suitable combination of academic and practical skills required for classroom teaching (https://www.dese.gov.au/teaching-and-school-leadership/teacher-education-ministerial-advisory-group).

TEACHER EDUCATOR Usually a qualified teacher and university lecturer, with postgraduate qualifications, involved in the education of preservice teachers.

TEACHING PERFORMANCE ASSESSMENT (TPA) A summative culminating assessment used to assess the practical skills and knowledge of preservice teachers. A TPA is a requirement for graduation and must be completed successfully in the final year practicum of a preservice teacher's ITE course.

THRESHOLD STANDARD This is the minimum expectation of performance to be conferred an overall pass (meeting the standard).

VALIDITY Validity refers to "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" and is collected from "an analysis of the relationship between the content of a test and the construct it is intended to measure" (AERA/APA/NCME, 2014, p. 14).

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Australian Institute for Teaching and School Leadership (AITSL). (2011; revised 2018). *Australian professional standards for teachers*. https://www.aitsl.edu.au/teach/standards

Australian Institute for Teaching and School Leadership (AITSL). (2015; revised 2019). *Accreditation of initial teacher education programs in Australia: Standards and Procedures*. https://www.aitsl.edu.au/docs/default-source/national-policy-framework/accreditation-of-initial-teacher-education-programs-in-australia.pdf?sfvrsn=e87cff3c_28

Department of Education, Skills and Employment. (DESE). (n.d.). *The Teacher Education Ministerial Advisory Group*. https://www.dese.gov.au/teaching-and-school-leadership/teacher-education-ministerial-advisory-group

edTPA. (n.d.). About. *edTPA*. http://www.edtpa.com/PageView.aspx?f=GEN_About EdTPA.html

Graduate Teacher Performance Assessment (GTPA). (n.d.). *Discover. GTPA*. https://www.graduatetpa.com/discover/.

Harlen, W. (2004, 29 November). Can assessment by teachers be a dependable option for summative purposes? [Paper presentation] at General Teaching Council for England Conference, New Relationships: Teaching, Learning and Accountability, London. In C. Adams & K. Baker (Eds.), *Perspectives on pupil assessment* (pp. 24–30). https://dera.ioe.ac.uk/14022/1/1104_Perspectives_on_Pupil_Assessment._New_Relationships__Teaching,_Learning_and_Accountability.pdf

Stanford Center for Assessment, Learning and Equity (SCALE). (n.d.). *Performance Assessment for California Teachers* (PACT). https://scale.stanford.edu/teaching/pact

Wiliam, D. (1994). Reconceptualising validity, dependability and reliability for national curriculum assessment. In D. Hutchison & I. Schagen (Eds.), *How reliable is national curriculum assessment?* (pp. 11–34). National Foundation for Educational Research. https://www.nfer.ac.uk/media/1422/91098.pdf

INDEX

Note: Page locators in **bold** refer to tables and page locators in *italic* refer to figures.