EXAMINATION STANDARDS

How measures and meanings differ around the world

Edited by Jo-Anne Baird, Tina Isaacs, Dennis Opposs and Lena Gray

Examination standards

'Toto, I've a feeling we're not in Kansas anymore.'
(Dorothy in *The Wizard of Oz*)

Examination standards

How measures and meanings differ around the world

Edited by Jo-Anne Baird, Tina Isaacs, Dennis Opposs and Lena Gray

UCL
IOE Press

First published in 2018 by the UCL Institute of Education Press, University College London, 20 Bedford Way, London WC1H 0AL

www.ucl-ioe-press.com

British Library Cataloguing in Publication Data:
A catalogue record for this publication is available from the British Library

ISBNs
978-1-78277-260-6 paperback
978-1-78277-261-3 PDF eBook
978-1-78277-262-0 ePub eBook
978-1-78277-263-7 Kindle eBook

Typeset by Quadrant Infotech (India) Pvt Ltd
Printed by CPI Group (UK) Ltd, Croydon, CR0 4YY

Cover image © Rawpixel Ltd/Alamy Stock Photo

Contents

List of figures

List of tables

About the contributors

Jo-Anne Baird is Director of the Department of Education at Oxford University and a member of the Oxford University Centre for Educational Assessment (OUCEA). Before coming to Oxford, Jo-Anne held academic posts at the Institute of Education, University of London and the University of Bristol. She was Head of Research at the Assessment and Qualifications Alliance (AQA), where she managed the research programme and was responsible for the standard setting systems for public examinations. Her first degree and doctorate were in Psychology and she has an MBA. Jo-Anne conducts a lot of work with government and industry partners, including acting as the Standing Adviser to the House of Commons Education Select Committee, a member of Ofqual's Standards Advisory Group and membership of the Welsh Government's Curriculum and Assessment Group. She is a member of the Editorial Board of the *Oxford Review of Education* journal and the International Advisory Board of *Assessment in Education: Principles, policy & practice* and is a former President of the Association for Educational Assessment – Europe.

Tina Isaacs began her career teaching university level history after receiving her doctorate from the University of Rochester, New York. In 1990 Tina joined the National Center on Education and the Economy, Rochester, New York, doing education policy research and development in curriculum and assessment. She moved to England in 1994 joining the National Council for Vocational Qualifications (NCVQ), and in 1997 the Qualifications and Curriculum Authority (QCA). She oversaw developments of General National Vocational Qualifications (GNVQs). In 1999 her responsibilities shifted from vocationally related to general qualifications, taking on responsibility for General Certificate of Secondary Education (GCSE), A levels and Advanced Extension Awards (AEA). She then led various teams before joining the Office of Qualifications and Examinations Regulation (Ofqual) as Head of 14–19 Regulation in 2007. Tina returned to higher education (UCL Institute of Education, London) in 2009. She is currently an honorary associate professor in educational assessment. Before that she was Programme Director for the MA in Educational Assessment, Principal Investigator of the Aligned Instructional Systems Project and a Co-Director at the Institute's Centre for Post-14 Research. She is currently

on the editorial board of *Assessment in Education: Principles, policy & practice*.

Dennis Opposs works at the Office of Qualifications and Examinations Regulation (Ofqual). Ofqual is the examinations regulator in England, and his current post is Standards Chair. He acts as an in-house expert on qualifications, standards and assessments. His present work focuses on standards and standard setting in GCSEs and A levels. In his early career, Dennis taught chemistry and other sciences in comprehensive schools in London and Hertfordshire. He worked in regulatory organizations in the 1990s including for the Qualifications and Curriculum Authority (QCA). His early responsibilities included GCSEs and A levels in science subjects and the initial development of National Curriculum science tests. In the last decade, Dennis has headed up Ofqual's Standards and Research Directorate and been in charge of the organization's reliability programme. He also led the development of standard setting of reformed GCSEs using a new grading scale and directed work on the comparability of standards between subjects. He is presently the representative for the Europe region on the Board of Trustees of the International Association for Educational Assessment (IAEA).

Lena Gray is Director of Research at the Centre for Educational Research and Practice (CERP) at AQA, England's largest GCSE and A level exam board, where she has worked since July 2014. Her role involves taking the lead in planning, implementing and evaluating a broad research programme that supports AQA's corporate aims. The work includes collaboration with researchers outside AQA, strategically engaging external stakeholders to maximize the impact of CERP's work on the education and qualifications system. Before joining CERP, Lena was Head of Policy, Assessment, Statistics and Standards at SQA, where she was responsible for the organization's research, establishing policy and introducing innovations to qualifications and assessment design and standard setting processes. She oversaw SQA's monitoring standards over time programme. Lena also played a major role in the government's review and development of curriculum and assessment 3–18. Prior to that, Lena held a number of positions in the SQA and predecessor organizations, chiefly related to qualifications and assessment reform. Lena started her career as a secondary teacher of English, before completing a doctorate in feminist literary theory

at the University of Strathclyde, where she also gained experience teaching and leading programmes of higher education.

Natia Andguladze is Associate Professor at Ilia State University School of Education, Georgia. She is also affiliated with the National Assessment and Examinations Centre (NAEC). At NAEC, Natia works on national and international assessments and other research projects in the area of teacher appraisal and school evaluation. Her research interests are concentrated around the application of educational assessment in education governance and administration, and cultural and structural aspects of equity in education. During the last ten years, Natia has co-authored various national education sector policy review projects and participated in national education sector and sub-sector strategy development initiatives together with colleagues from the Ministry of Education and Science and international and bilateral agencies.

Eva L. Baker is a Distinguished Professor of Education at UCLA and Founding Director of the National Center for Research on Evaluation, Standards and Student Testing (CRESST). Supported by foundations, governments and business, CRESST's R&D focuses on assessment and technology, evaluation and learning. Eva initiated and served as President of the World Education Research Association (WERA), following her tenure as President of both the American Educational Research Association (AERA) and the Educational Psychology Division of the American Psychological Association (APA). She has received numerous awards, including both AERA's 2014 Robert L. Linn Lecture and the 2013 E.F. Lindquist Awards.

Steven Bakker is an international consultant for educational assessment. Before he started his own business, Steven held positions at CITO, the Dutch National Institute for Educational Measurement and at ETS (Educational Testing Services, Princeton, US) as Executive Director of ETS Europe. He is the founder and past-president of AEA–Europe.

Betsy Brown Ruzzi is Vice-President of the National Center on Education and the Economy (NCEE) and Director of its Center on International Education Benchmarking (CIEB). During her career at

the National Center she helped create the National Institute for School Leadership, the National Skill Standards Board, the Commission on the Skills of the American Workforce and the National Board for Professional Teaching Standards. She serves on the Board of Governors of the Academy of Education Arts and Sciences and is a member of the Comparative International Education Society. In addition to working at the National Center, Betsy worked in the United States Congress, in the British Parliament and in the Governor's Office in Massachusetts.

Matthew Campbell was until recently Principal Research Officer at the Queensland Curriculum and Assessment Authority, the statutory body responsible for providing curriculum, assessment, certification and tertiary entrance services to Queensland schools. He is currently Senior Lecturer in the Learning and Teaching Unit at the Queensland University of Technology. Matthew also holds an honorary adjunct lecturer position with Griffith University in the Centre for Learning Futures at Griffith University. He has more than 20 years' experience in a variety of roles in the secondary and tertiary education sectors, including academic positions at several Australian universities. Matthew has taught in the areas of curriculum, assessment, science education and professional ethics. His previous research has largely focused on the development of professional identity and work-based learning.

Alina von Davier is Senior Vice-President at ATC Inc. and leads the company's research and development group, ACTNext. She is also Adjunct Professor at Fordham University, New York. Alina is a pioneer in the development and application of computational psychometrics, conducting research on blending machine learning algorithms with psychometric theory. Her co-edited volume on *Computerized Multistage Testing* (2016) and edited volume on *Statistical Models for Test Equating, Scaling, and Linking* (2013) were selected as the winners of the Division D Significant Contribution to Educational Measurement and Research Methodology award at AERA. As Principal Investigator, she has received funding from the National Science Foundation, the Spencer Foundation, the MacArthur Foundation, US Army Medical Research and the Army Research Institute. She serves as Associate Editor for *Psychometrika* and the *Journal of Educational Measurement*.

Gudrun Erickson is Senior Professor of Education in language and assessment at the University of Gothenburg, Sweden, with extensive experience of national and international projects on assessment. She is also a scientific advisor for the development of national assessment materials for foreign languages; previously she was the project leader. Her research focus is on collaboration in test development.

María Teresa Flórez Petour is Assistant Professor in the Pedagogical Studies Department of the University of Chile, where she also coordinates the Centre for Studies on Educational Assessment. She is Research Associate of the Oxford University Centre for Educational Assessment. Her research interests include: assessment policies from a complex, critical and historical perspective; validity in high stakes assessment systems; implementation of Assessment for Learning. She has been involved for several years in professional development programmes, consultancy work and research around these topics both in Chile and the UK.

Roger-François Gauthier is a scholar in Education Science (PhD, University Lyon 2) and a civil servant who has a long involvement in French educational policy. He focuses on the issue of defining school knowledge in contemporary societies. As a general inspector for education in France, he has reported on most common issues of education management, including aspects of curriculum, system evaluation and student assessment. He teaches educational policies at Paris-Descartes La Sorbonne University, basing his approach and teaching methodology mainly on methods related to comparative education. He is currently a member of the French High Council for Curriculum (CSP). He has been consulted as an expert by many governments, national and international organizations, such as UNESCO. He has published books, articles and reports in France and various countries and in international organizations, including UNESCO and OIF.

Francisco Javier Gil Llambías is Director of the Programme of Inclusive Access Equity and Permanence of the Universidad de Santiago de Chile and Director of the Chair on Inclusion in Higher Education of UNESCO. He has a PhD in Chemical Sciences and is a deacon of the Catholic Church. He is co-author of about 60 ISI publications; the *Report*

of the University Reconciliation Commission of the Universidad de Santiago de Chile (USACH) (1990); and contributed to four instruments to increase inclusion in higher education. He has previously acted as Rector of the Universidad Católica Silva Henríquez and Academic Vice Chancellor and Dean of the USACH.

Kristine Gorgen is a DPhil (PhD) student at the Oxford University Department of Education and a research assistant at the Oxford University Centre for Educational Assessment (OUCEA). After receiving BAs from the Institut d'études politiques de Paris au Havre (Sciences Po) and Columbia University in New York, Kristine came to Oxford to read for the MSc in Comparative and International Education. She was awarded a distinction for her thesis on the impact of PISA on social justice discourses in German policymaking. Since April 2016 Kristine has been a research assistant at OUCEA, where she was on the organizing committee for the 2016 PISA Conference, supports the ESRC-funded project Assessment for Learning in Africa and has worked on the standard setting and maintaining in national examinations project. Kristine is the recipient of the Philosophy of Education Society Great Britain (PESGB) Doctoral Studentship. In the autumn of 2017 Kristine was a visiting researcher at the Social Sciences Centre Berlin (WZB).

Jan-Eric Gustafsson is Emeritus Professor of Education at the University of Gothenburg, Sweden. His research is primarily focused on basic and applied topics within the field of educational psychology, and particularly on models for the structure of cognitive abilities, and on assessment and development of abilities, knowledge and skills.

Sarah Howie is Director of the Africa Centre for Scholarship and Professor at Stellenbosch University, South Africa. She is Deputy Chair of the Board of South African Qualifications Authority and was a member of the Umalusi Assessment and Standards Committee and Ministerial Committee in 2013/2014 on the National Senior Certificate. She was National Research Coordinator of the International Association for the Evaluation of Educational Achievement (IEA)'s TIMSS, SITES and PIRLS studies for South Africa.

Áine Hyland is Emeritus Professor of Education and former Vice-President of University College, Cork, Ireland. She was a member of Ireland's first Curriculum and Examinations Board in the 1980s and was Vice-President of the Irish Research Council for the Humanities and Social Sciences from 2006 to 2012. She has written articles and academic papers on curriculum and assessment from historical, policy, practice and comparative perspectives.

Jean-Pierre Jeantheau is National Project Leader in the French National Agency for Fighting Illiteracy. He is affiliated with University Lyon 2 for the training of students in statistics and previously taught international surveys methodology in the Catholic University of Paris. He holds a PhD degree in Sociolinguistics from Paris-Sorbonne University and a Master's Degree in Education.

Sandra Johnson is an assessment consultant, with extensive experience in teaching, research and assessment practice (large-scale assessment and national qualifications). Her publications include books, attainment survey reports, teaching resources, qualification evaluation reports and academic journal articles. She is a member of the editorial boards of *Assessment in Education: Principles, policy & practice* and *Educational and Psychological Measurement*.

Anil Kanjee is Research Professor and Coordinator of the Postgraduate and Research Programme in the School of Education, Tshwane University of Technology. He also serves as a Research Fellow at the Centre for International Teacher Education (Cape Peninsula University of Technology) and the Oxford University Centre for Educational Assessment. His research focuses on addressing the challenge of equity and quality in education, paying particular attention to the use of large-scale and classroom assessment for improving learning and teaching, learning and learner voice in schools and teacher professional development.

Anna Lind Pantzare is an upper secondary school teacher in mathematics and physics and has been working with the Swedish national

tests in mathematics and science for nearly 20 years – from the beginning as a test developer and since 2013 as a project manager. Anna is currently working on her PhD in Educational Measurement at the Department of Applied Educational Science, Umeå University. Anna's research focuses on aspects of validity and comparability in national testing, and she has a special interest and expertise in standard setting. Anna has recently published a chapter in *Standard Setting in Education: The Nordic countries in an international perspective* (2017, Springer).

Graham S. Maxwell is currently Honorary Professor in the Learning Sciences Institute Australia (LSIA) of the Australian Catholic University (ACU). He specializes in policy and practice in educational assessment, with a focus on standards and moderation. Initially a high school teacher of mathematics and science, he spent 30 years in the School of Education at the University of Queensland, followed by a three-year appointment as Deputy Director of the Queensland Studies Authority (QSA), predecessor of the Queensland Curriculum and Assessment Authority (QCAA).

Joshua McGrane is Deputy Director and Senior Research Fellow at the Oxford University Centre for Educational Assessment (OUCEA) who specializes in psychometric research. He previously held positions as a psychometrician at the New South Wales Department of Education and as a postdoctoral researcher at the University of Western Australia.

Hugh McManus is Assistant Head of Examinations and Assessment at the State Examinations Commission, the state agency responsible for running the second-level public examinations in Ireland. Hugh heads the Commission's research unit and also oversees the work of many of the subject specialists responsible for the Commission's examinations. He represents the Commission on the Transitions Research Group – an inter-agency working group overseeing research related to the Irish government's Transitions Reform agenda. These reforms are intended to improve the transition from second- to third-level education in Ireland. Hugh also represents the Commission in its membership of AEA–Europe and the IAEA. Prior to his current appointment, he managed all of the Irish Leaving Certificate mathematics examinations, and this included managing these examinations through the implementation period of Project Maths,

a substantial national programme of reform of mathematics education in Ireland. Before joining the Commission in 2003, he spent five years as a post-primary schools inspector, a period that included a brief secondment to the World Bank as an Education Specialist, and before that he spent eight years as a mathematics teacher in a large secondary school.

Iwa Mindadze is Deputy Director of the National Assessment and Examinations Centre (NAEC). NAEC, the largest professional organization acting in the field of assessment in Georgia, conducts different high-stake large-scale qualifying and certification exams. Aside from this Iwa is involved in many projects across the country, delivering technical assistance, training and advice relating to educational assessment. At the same time Iwa is Associate Professor at Ivane Javakhishvili Tbilisi State University, faculty of Psychology and Education. In recent years Iwa has been teaching at the Goethe Institut, Tbilisi, and also acting as an international consultant in the field of educational assessment in countries such as the Ukraine, Kyrgyzstan and Kosovo. He has been Visiting Professor at the University of Saarland, Germany. His research interests are focused on Psycholinguistics, and the relationships between language and culture, and language and mind.

Deanna L. Morgan is Senior Director of Psychometrics at the College Board, US. She has worked in the assessment and education community for 24 years, seven of those as a classroom teacher and 17 as a psychometrician. She earned her Doctorate in Educational Psychology with a focus on Research, Evaluation, Measurement, and Statistics from the University of Georgia in 2001. Her research interests include standard setting, generalizability theory, reader reliability, large-scale assessment, college placement testing and the assessment of students with disabilities. She leads a team of psychometricians with responsibility for the Advanced Placement programme and state contracts for the use of the SAT for accountability purposes under the Every Student Succeeds Act (ESSA). Among her many publications are two chapters in *Handbook on Measurement, Assessment, and Evaluation in Higher Education*, edited by C. Secolsky and D.B. Denison (2011).

Michael O'Leary holds the Prometric Chair in Assessment at Dublin City University and is Director of the Centre for Assessment Research Policy

and Practice (CARPE) at the Institute of Education, Dublin City University. He leads a programme of research at CARPE focused on assessment in education and in the workplace. Current projects include teachers' attitudes to and use of standardized tests, learning portfolios in higher education, the assessment of critical thinking, the use of animations to assess teachers' tacit knowledge, state of the art in technology-based assessment and the use of situational judgement tests to measure soft skills.

Alejandra Osses is Head of the Development and Analysis Unit at the Department for Education Measurement, Assessment and Registry (DEMRE), University of Chile. She is a sociologist and holds a PhD in Education with research and work experience in the area of education assessment and policy. Her work experience includes the Centre for Advanced Research in Education at the University of Chile – where she currently participates as Associate Researcher – the Assessment Research Centre (ARC) at the University of Melbourne and the Australian Council for Educational Research (ACER). Currently, she works in the Department for Measurement, Assessment and Education Registry at the University of Chile, leading the team responsible for developing new assessment instruments for tertiary education admissions and for analysing the results of the current university entry test, the PSU.

Mary Richardson is Associate Professor in Education and Programme Leader for the MA in Educational Assessment at the UCL Institute of Education. Originally a theatre and education practitioner, Mary worked with children in early years and schools and then moved into educational research in both policymaking and academic settings.

Emmanuel Sibanda is a statistician and an education researcher. He has more than 20 years of research and data analysis expertise, spanning three universities and a Quality Council. Emmanuel is currently Executive Manager of Qualifications and Research branch of Umalusi, an exam board and regulator. He has managed more than 30 research projects of various sizes and complexities at Umalusi, which resulted in published and unpublished reports. Emmanuel holds a Master's Degree in Mathematical Technology and a BSc (Hons) in Mathematical Statistics. He is currently studying towards a PhD in Mathematics Education. He has read more

than 20 research papers at national and international conferences and co-published two research papers in accredited journals.

Gordon Stobart is Emeritus Professor of Education at the UCL Institute of Education and Honorary Research Fellow at Oxford University. Having worked as a secondary school teacher and an educational psychologist, he spent 20 years as a senior policy researcher, first as head of research at an examination board, then at government education agencies. His books include *The Expert Learner: Challenging the myth of ability* (2014, OUP/McGraw-Hill) and *Testing Times: The uses and abuses of assessment* (2008, Routledge).

Rachel Taylor is Research Fellow (Standards) at the regulatory body, Ofqual. As part of her role she is involved in technical discussions with exam boards relating to the maintenance of GCSE and A level standards, and in monitoring exam boards' awarding outcomes each exam series. She has also been involved in a number of research projects relating to the maintenance of standards. In early 2018 she completed a DPhil in Education at the Oxford University Centre for Educational Assessment. Her research focused on the implications of early and multiple entry in GCSE Mathematics for maintaining examination standards, using a mixed methods approach.

Peter Tymms is Director of the international study on Performance Indicators in Primary Schools (iPIPS) at Durham University and was Head of Department in the School of Education at Durham University until 2013. Before that he was Director of the Centre for Evaluation and Monitoring (CEM), which runs projects monitoring millions of pupils across the UK and beyond each year. He set up the PIPS project, which has assessed more than three million children worldwide. He has published more than 100 scientific papers and brought in millions of pounds worth of research grants.

María Leonor Varas has been Director of the Department for Education Measurement, Assessment and Registry (DEMRE) at the University of Chile since 2015. She is also Professor in the Department of Mathematical Engineering and Associate Researcher at the Centre for Advanced Research in Education, both at the University of Chile. Prior

to 2015, she worked at the Centre for Mathematical Modelling, at the same university. Her preparation includes a Doctorate in Engineering Science specializing in Mathematical Modelling (University of Chile) and a professional degree in Mathematical Engineering (University of Chile). Her research experience covers mathematics, mathematical education and educational assessment.

Christina Wikström is Associate Professor at the Department of Applied Educational Science, Umeå University, Sweden. Christina's current research is mainly focused on issues relating to admission to higher education, and in particular on validity issues relating to the use of school grades and tests as instruments for selection. Christina is Head of the doctoral programme in Educational Measurement at Umeå University and teaches courses in assessment on undergraduate and graduate level. She is also one of the coordinators for a national research school in quantitative research methods in education (QRM), a member of the Editorial Advisory Board for the journal *Assessment in Education: Principles, policy & practice* and an associate editor for the Frontiers' Journal *Assessment, Testing and Applied Measurement*.

Preface

Alina von Davier

The education of the next generation has always been important in all cultures, although a formal educational system was developed or adopted only sporadically in antiquity. One of the earliest formal schools was developed in Egypt's Middle Kingdom, for example. Move forward 2,000 years, about half way between the Middle Kingdom and the present day, and we encounter some of the basic concepts that shaped Western education: the dialectic method of Socrates and the didactic method of Plato. In particular, Plato's recommendations for a national educational system outlined in *The Republic* (380 BC) became one of the most influential volumes for the education of many generations. Throughout *The Republic*, Plato returns to the concept of equity in education. Although I am writing this Preface 2,000 years after Plato, equity still occupies a central role when we talk about educational standards and goals for educating the young and preparing them to be the standard bearers who will carry human society forward. National examinations and consistent methods of setting standards are one way of approaching this goal.

Policymakers and education experts have always been interested in fair and accurate ways of designing and comparing educational systems and students' achievement. This volume, *Examination Standards: How measures and meanings differ around the world*, continues this tradition and looks at how the method of setting educational standards varies across different countries (Chapters 1–4), and then provides a detailed exploration of the standards for education in several nations. I absolutely concur with the editors' perspective on this overview of educational standards around the world in the first chapter that

> we are challenging the notion that there is a single (superior) way of thinking about national examinations. […] psychometrics is the dominant paradigm and the one most frequently put forward as most advanced, technically sound and theoretically robust. To stake a claim for one of the paradigms being superior requires a treatment of its utility in relation to purpose and the values of its users (p. 5).

Consequently, the meaning of standard setting in a psychometric context is of secondary importance for this volume, although Chapter 4 provides an overview of the psychometric arsenal for test score comparability across different test forms (equating), or across different tests (concordance), standard setting panels and methodologies for establishing achievement levels and other approaches.

After the establishment of the European Union, all European countries faced the dilemma of selecting and accepting students from all over the EU without a formal alignment across educational systems. Quite naturally, questions arose regarding the methods for setting comparable educational standards: either those based on assessments, or on other types of evidence of educational experience.

On the other side of the Atlantic, in the US, there is neither a common curriculum nor a centralized assessment system; hence, different tools for comparability and standards were developed, from the Advanced Placement (AP) examinations to the National Assessment of Educational Progress (NAEP) survey assessment. The generation of the Common Core State Standards (for mathematics and English) and recently the Next Generation State Standards (for science) was motivated by the need to make sense of the differing educational criteria required by individual states. This effort was followed with standard large-scale assessment programmes such as Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (SBAC).

Internationally, the interest in comparing countries' educational standards across cultures led to large-scale survey assessments like the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS). Moreover, with digital learning and assessment systems (LASs) and freely available Massive Open On-line Courses (MOOCs), access to education (the quality of which is in open debate) is now available to more people around the world, regardless of their location. Returning to the topic of equity, as students around the globe pursue learning through these evolving digital means, those of us working in the field of education must endeavour to identify meaningful standards and articulate the value of achievements like the 'levels' and 'badges' earned in this new sphere.

Situated in this geo-historic context, I welcome this new volume on setting educational assessment standards across many countries. The volume provides practitioners and policymakers with a fresh and timely mosaic of information on different countries' systems that can be used for inspiration or comparison. The volume first discusses the common methodologies for

establishing levels of scholarly achievement and for comparing test scores, and subsequently illustrates individual experiences in different nations. The editors invited experts to comment and reflect on each of these country-focused chapters, which provide a multifaceted perspective of each system.

It is interesting to read about the diversity of challenges that different countries struggle with, despite an appearance of common goals for educating youth. The influence of each country's unique political and geo-economic history, the levels of heterogeneity within populations and the (de)centralization of educational and political systems influence how new generations are taught and how they will fare in comparison to their peers across the world.

Most recently, some of us have been working on blending learning and assessment in ways that would allow students to demonstrate their mastery of skills and knowledge as part of a (digital) learning and assessment system. This volume brings back the question of comparing different learning and assessment systems. How would we do that? Can machine learning help achieve the alignment of learning and assessment systems using crosswalks of taxonomies and achievement levels as von Davier *et al.* (2017) propose? Or, more generally, do we expect educational standards to change in the age of Artificial Intelligence (AI)? AI and the digital revolution have disrupted nearly every aspect of modern society. Is there a compelling reason to expect educational standards will be exempt?

Another line of current research is focused on developing appropriate methodologies for assessing hard-to-measure twenty-first-century skills like collaborative problem solving and creative thinking. PISA, for example, became a leader in experimenting with the measurement of innovative domains across multiple countries and cultures at scale. How will these measurements transfer to local educational systems? Will we use the PISA assessment to establish achievement levels for collaborative problem solving at the individual level in different educational systems, for example? Often, once a skill is on the international tests, it will penetrate the curriculum in individual countries. This points to the tremendous responsibility of international testing. To return to Socrates's theories and then to his own destiny, society's perceptions are not always very forgiving, so the stakes are high when it comes to educational standards. Getting it right is important.

A good book always leads to questions and research ideas. After reading this volume I have plenty of both. I encourage readers to study it thoroughly and use the rich information contained herein to build and develop their own questions and research agendas.

References

von Davier, A.A., Yudelson, M. and Blum, A. (2017) 'Thoughts on the role of test comparability in learning and assessment systems'. Presentation delivered at the Iowa Equating Summit, Iowa City, Iowa, September.

Plato (380 BC) *The Republic*. Online. www.idph.net/conteudos/ebooks/republic.pdf (accessed 7 June 2018).

Acknowledgements

The editors would like to gratefully acknowledge the contributions of a range of people to this book. First, we would like to acknowledge the attendees of the international symposium, held at Brasenose College at the University of Oxford, on 28–30 March 2018. A list of attendees can be found in Appendix A. Central to the project's aim was open dialogue with knowledgeable colleagues from a wide range of jurisdictions. Attendees brought expertise from an even broader range of settings than their employing institutions may suggest, due to their professional experiences in wider jurisdictions. This added a depth and richness to the proceedings that was of huge benefit. Robust exchanges were anticipated and experienced, and we are grateful for the challenges posed. Second, we would like to thank Kate Kelly from AQA for her contribution as Research Assistant to the project. Third, we thank Eleanor Gaspar, Joanne Hazell and Kristine Gorgen (OUCEA) and Lindsay Simmonds (AQA) for their assistance with proofreading and for saving us from some of the inconsistencies and errors that existed in the early texts. Fourth, the project would not have been possible without funding from AQA, Ofqual and the ESRC [Grant no 1609-VPRC-248]. We are grateful to the leaders of those organizations for their vision in engaging with and supporting an international project of this kind. Fifth, we are grateful to our editor, Nicky Platt, at the UCL IOE Press, for supporting us and giving us the scope to shape the book. Sixth, we are grateful to members of AQA's Research Committee and Ofqual's Standards Advisory Group for helpful discussions on the research methods to be deployed in the project. Finally, the book's case study authors and associated commentary authors have brought the issues to life, and the project would have been much less vivid without the important issues that they depict.

Part One

Researching national examination standards

1

Chapter 1

The Standard Setting Project: Assessment paradigms

Jo-Anne Baird and Dennis Opposs

Examination standards, what they mean and how they are measured, are often assumed to be unproblematic. As this book shows, however, very different approaches to defining and measuring standards are used around the world. To understand our own practices, which we often take for granted, we need to compare them with how examination standards are thought about at other times and in other places. By applying historical and comparative lenses to our practices, we can begin to classify and codify a field that is currently highly fragmented. In this chapter we outline three distinctive approaches to thinking about educational assessment in general. First, we trace the history of educational and psychological assessment. As part of that history, we then refer to an international project conducted in the 1930s, the aim of which was to advance the science of educational assessment. Finally, we introduce the project that generated this book and outline its remaining chapters.

The applied fields of assessment

The history of testing can be traced back well beyond that of intelligence testing. Imperial examinations, used for entry to the civil service in China, were first created in 124 BC (Roberts, 2006: 31). Aspects of current national school leaving examinations bear a great deal of similarity to those approaches, with students sitting written examinations in invigilated conditions and the results being used for selection purposes. With the invention of intelligence testing (Alfred Binet) and subsequent developments in psychometric testing, alternative ways of thinking about assessment became available. Psychometrics, first used in an educational setting, generated a great deal of interest in the relationship between intelligence and examination results. However, its scope grew beyond education to encompass psychological factors such as personality and psychological disorders. Additionally, the field of occupational psychology grew, with its own requirements. Thus, we can distinguish three applied areas for assessment: educational, psychological and workplace.

Table 1.1: Four editions of *Educational Measurement*

Editor Year	Standard setting	Educational	Psychological	Workplace
Lindquist 1951	Flanagan. Chapter 17 – Units, scales and norms	e.g. Tyler. Chapter 2 – The functions of measurement in improving instruction	Darley & Anderson. Chapter 3 – The functions of measurement in counseling	Ryans & Frederikson. Chapter 12 – Performance tests of educational achievement
Thorndike 1971	Angoff. Chapter 15 – Scales, norms and equivalent scores	e.g. Krathwohl & Payne. Chapter 2 – Defining and assessing educational objectives	Davis. Chapter 18 – Use of measurement in student planning and instruction	Fitzpatrick & Morrison. Chapter 9 – Performance and product evaluation
Linn 1989	Petersen *et al*. Chapter 15 – Scaling, norming and equating	e.g. Nitko. Chapter 12 – Designing tests that are integrated with instruction	Shepard. Chapter 17 – Identification of mild handicaps	Jaeger. Chapter 14 – Certification of student competence
Brennan 2006	• Kolen. Chapter 5 – Scaling and norming • Holland and Dorans. Chapter 6 – Linking and equating • Hambleton and Pitoniak. Chapter 15 – Setting performance standards	e.g. Shepard. Chapter 17 – Classroom assessment	–	Clauser *et al*. Chapter 20 – Testing for licensure and certification in the professions

An historical snapshot of the field of educational assessment can be gleaned from the landmark publications of the US book, *Educational Measurement*, which has been published in four editions, spanning 55 years (Table 1.1). We see that there are chapters reflecting standard setting in each edition. With American authors, the underlying paradigm of the book is psychometric. As would be anticipated from the title, many of the chapters reflect educational concerns. Each edition recognizes the relevance of performance assessment, separate chapters on this topic appearing in the first three editions. In the fourth edition, however, the chapter on this topic is broader, referring to licensure and certification. That there are different chapters on educational, psychological and workplace matters could be seen as thematic, with each of these areas of application producing particular challenges for the field, just as the specific theme of standard setting does. The chapters in *Educational Measurement* all reside within the psychometrics paradigm, even if they have applications in different fields. However, as argued below, quite distinctive ways of thinking about educational assessment have arisen that suggest and address different questions and ways of interpreting the evidence that is collected about them.

Over the timespan of the publication of these editions, there were developments in the applied fields of educational, psychological and workplace assessment, but they were not always entirely compatible developments. We outline three different paradigms of assessment that have been developed. In Thomas Kuhn's (1962) terms, a paradigm offers a 'universally recognized scientific achievement that, for a time, provides model problems and solutions for a community of researchers' (p. 311). A paradigm is a guide to what is to be observed and studied (below termed the attribute), what kinds of questions we might ask, how the questions should be structured and how the results of this investigation should be interpreted. In assessment, this has come to be represented through the technologies that are used such as the assessment formats and the analysis techniques. Distinguishing the paradigms provides a framework for the field and explains some of the tensions that arise in examination systems. The paradigms are distinctive traditions of assessment, involving different assumptions and philosophical underpinnings. They are outlined below as idealized types so that they can be clearly differentiated. We also note that although Kuhn's definition of paradigm refers to a universally held model, we are outlining three paradigms that are in competition. In Kuhn's terms this is a pre-paradigmatic phase. However, Kuhn's work focused on the hard sciences rather than the social sciences. Worldviews are much more

contested in the social sciences due to their social and political contexts, though climatologists may disagree with this characterization. Instead of casting educational assessment as being in a pre-paradigmatic phase, we think it more helpful to think of the paradigms as positions in the field.

Assessment paradigms

In suggesting that there is more than one way of thinking about assessment, we are trying to describe both the history and state of the art of national examinations. Additionally, we are challenging the notion that there is a single (superior) way of thinking about national examinations. As previously discussed, psychometrics is the dominant paradigm and the one most frequently put forward as most advanced, technically sound and theoretically robust. To stake a claim for one of the paradigms being superior requires a treatment of its utility in relation to purpose and the values of its users. Our position is that multiple perspectives exist because psychometrics has not adequately addressed the purposes and values that are prioritized in some national contexts; that is one of the messages of the International Examinations Inquiry, described below, that we believe persists. We are not arguing that it will always be so, merely that this is the current position. With the homogenizing influences of international testing organizations such as the Organisation for Economic Co-operation and Development (OECD) and the International Association for the Evaluation of Educational Achievement (IEA), we might anticipate the spread of the psychometrics paradigm (see Grek, 2009). Indeed, there is evidence of such effects already (Baird *et al.*, 2016).

We are not the first to use the term 'paradigm' in relation to assessment. Andrich (2004) discussed different paradigms at work within the field of psychometrics. He contrasted approaches that sought to model psychological phenomena (e.g. two-parameter models) and those that tried to measure (one-parameter, Rasch models). In Andrich's terms, measurement is conducted on an interval scale (see below) and therefore the Rasch model is necessary. Additionally, Andrich (2004) argued that we ought to design tests so that the data fit the Rasch statistical model because otherwise we did not have an interval scale and therefore were not measuring at all. This has been a very heated debate in the field. The distinction between modelling and measurement is important, but all of our paradigms set out to do more than model data. This leaves us in the tricky territory of what it means to measure and whether that necessarily entails an interval scale, leading us to the debates around measurement scales discussed in

this chapter. As will be seen below, our paradigms take different positions on this, but principles often seem to melt in the face of pragmatics and heroically complex calculations are conducted on examination results that we would recognize are not warranted if we stuck to our original positions on measurement scales.

A paradigm is an approach to assessment that has implications for practice such as test design, quality assurance, analysis, data interpretation and reporting techniques. Underlying the paradigms are different philosophical positions and notions of what it means to assess and what the results should look like. This includes, but is not limited to, measurement scales. People do not often reflect too much on their fundamental beliefs, so in practice we see few pure paradigmatic examples of educational assessment. Nonetheless, it is important to distinguish these paradigms to understand the field; its history, ways of operation and its tensions.

Psychometrics paradigm

Psychometrics is concerned with measurement of the mind and refers to a way of thinking about how tests should be constructed, administered, analysed and the outcomes interpreted. Its origins are in the psychological testing field, particularly intelligence testing. Sir Francis Galton, in a book entitled *Hereditary Genius* (1869), laid some of the conceptual groundwork for psychometrics, including scatterplots, which were the prelude to the formal development of statistical techniques of correlation. James McKeen Cattell coined the term 'mental test' and worked with Charles Spearman on the development of factor analysis, which can help describe multiple factors assessed in a test simultaneously. In educational assessment, it is more commonplace for a single factor to be considered; in fairness, multiple factors have been tricky to handle in most applications other than personality testing.

Intelligence testing grew at a time when psychology was trying to prove itself as a science. Statistics were developed contemporaneously to solve the kinds of issues that were being grappled with in the field of mental testing. Importantly, the field of measurement error was already well developed, having been tackled in astronomy for some time (Porter, 1986). There, it had been observed that individual measurements could contain error and that a set of measures followed a normal distribution, or bell curve, in which most of the measures were in the middle and fewer at the extremes (with more error). A wealth of statistical techniques was constructed based upon the properties of normally distributed data. This

proved incredibly valuable because normal distributions were observed for a range of biological and social phenomena, such as height, shoe size and other population characteristics.

An important leap made by Galton was to theorize that mental phenomena were also normally distributed in the population (see Goldstein, 2012). It followed that when tests were constructed, the scores should be normally distributed. Therefore, the construction of tests was designed to meet this principle. Although there is not space in this chapter to compare and contrast all of the features of the different paradigms, we will point to a few. Here, let us consider the necessity of normally distributed exam scores. If we create an examination and find that the scores are not normally distributed, does this mean that there is something wrong with our test? Should we change the questions? This is precisely the issue facing us if we sign up to a psychometrics paradigm. But if our purpose is simply to find out what children know about the biology curriculum in Year 10, then we might expect that results could have very differently shaped distributions. Also, if an important purpose of the test is to discriminate between children at particular points in the scale so that they can be graded, a normally distributed score could be a disadvantage. If a lot of children score around the mean, it might be hard to classify them without error at around the midpoint, which might be a very meaningful point of classification for the examination.

However, outcomes from psychometric tests are typically not graded. Outcomes are usually scores that are internal to the test itself and are not exchangeable across different kinds of tests in the way that letter grades are intended to be. We might well question this exchangeability – this is often done and there is a research literature on techniques for conducting such investigations (Newton *et al.*, 2007). Fundamentally, the attribute of interest in psychometric testing is an unobservable, latent trait, which can only be measured with error. These traits might be viewed as a fixed feature of an individual test-taker. A lot of emphasis is therefore placed upon the internal reliability of the test; whether the items all measure the same thing. In terms of validity, construct validity is foregrounded; whether the test measures this latent trait. A traditional test format in psychometrics is multiple choice. This need not necessarily be the case, and while statistical techniques have been developed to tackle a range of formats, multiple choice testing still dominates in this tradition. Tests are not always curriculum-related. After all, the latent traits might be viewed as fixed features of individuals so the curriculum would be something of an aside in this way of viewing things.

With its origins in a scientific approach, psychometric tests are conducted in controlled conditions so that users of the tests can be sure that the results were caused by the latent traits of the individuals taking the tests rather than other factors (such as the conditions of test taking). In the psychometrics paradigm, examination standards are set by subject matter experts in combination with psychometricians. Psychometric tests are operating well when they discriminate between the test-takers and rank-order them appropriately. Results might be used to distinguish a particular percentage to pass the test, percentages to be awarded each grade or a percentile score might be awarded. Norm-referencing is the prototypical approach to standard setting in this paradigm. This method involves deeming a certain proportion of the population as having passed or being graded at a specific level. Chapter 4 discusses norm-referencing more fully.

Critics of psychometrics might argue that it does not deserve to be labelled a paradigm and is instead merely a set of statistical techniques (Goldstein, 1979; Goldstein and Wood, 1989). However, the psychometrics tradition is more than a set of statistical techniques. Indeed, if it were only statistical models, we might not have seen the kinds of heated debates and paradigm wars that have occurred (see Chapter 14). Psychometrics is a way of construing the social world in educational assessment. Baird and Black (2013) argued that psychometrics looked like an answer to somebody else's problems when they outlined the implications of the use of psychometrics for examinations. For example, it is commonplace for national examinations to have transparent structures and content and for past papers to be published. In this way, teachers and students can see what needs to be learnt. However, it is important for the stability of statistical parameters of psychometric tests that the test items are kept secure. This may be too big a price tag if your main purpose is improvements in education. Individuals might use the statistical techniques of psychometrics without signing up to the philosophical or theoretical beliefs that belong with this paradigm. But it is our observation that this leads to difficulties of various kinds. Working across paradigms leaves practitioners without a consistent structure and leads to incompatible practices and ways of thinking about and explaining results. Practitioners have the options of tolerating this situation or producing a coherent narrative for their cross-paradigm working. In effect, muddled thinking and practice is frequently observed.

We lump together the one- and two-parameter models distinguished by Andrich (2004) as having different paradigms underlying them. Both approaches have similar underpinning beliefs, but they differ in terms of what should be done when the data do not fit the statistical model. Two

(or more) parameters are introduced in some models to make the model fit the examination data better. An alternative would be to say that the data are wrong and to produce tests that fit a one-parameter model better. While this is important, the purpose of this book is broader than this distinction: our position is that psychometrics as a field is attempting to measure psychological phenomena, even if some are dissatisfied with how some of the psychometrics community go about it.

Outcomes-based assessment paradigm

Distinct from psychometrics and arising from the occupational, workplace application of educational assessment is the outcomes-based paradigm. This approach has its disciplinary roots in management theory such as Taylorism (e.g. see Neumann, 1979), but it can be traced back further to the apprenticeship tradition. A boost for the promulgation of outcomes-based education came from US Office of Education state sponsorship for ten colleges to develop teacher training programmes (Tuxworth, 1989). Following this, federal funding was given for the development of vocational training programmes. Tensions over the role of knowledge and competence were hotly debated during these developments.

A central aspect of the outcomes-based assessment paradigm is the setting of criteria, goals, or outcomes. Tyler (1949), in his classic book entitled *Basic Principles of Curriculum and Instruction*, observed 'educational objectives become the criteria by which materials are selected, content is outlined, instructional procedures are developed and tests and examinations are prepared' (p. 3). Thus, although the setting of objectives is common in educational assessment, it is emphasized in this paradigm and plays a more central role in assessment.

The outcomes-based education movement underpins the outcomes-based assessment paradigm. Spady's (1977) work in the US outlined the theoretical basis for competency-based education, and it is this term that underpins the approach in general. In his terms, competency itself was an indicator of successful performance in life-role activities rather than discrete cognitive, manual, or social capacities. 'Measurement' in these terms would require considerable conceptual and technical development according to Spady (1977: 25). Some authors took him up on this challenge (e.g. Jessup, 1991). A definition of competency-based assessment is also helpful:

> Competence-based assessment is a form of assessment that is derived from the specification of a set of outcomes; that so clearly states both the outcomes – general and specific – that

> assessors, students and interested parties can all make reasonably objective judgments with respect to student achievement or non-achievement of these outcomes; and that certifies student progress on the basis of demonstrated achievement of these outcomes. Assessments are not tied to time served in formal educational settings (Wolf, 2001: 1).

Rather than a theoretical latent trait, in the outcomes-based tradition, the attribute of interest is competency in a specific set of skills and knowledge as demonstrated in performance. Usually, the purpose is to certify that a person is fit to practise in the occupation of interest, such as an electrician. Wolf (1995: 2) argued that three components of outcomes-based assessment were important:

1. an emphasis upon multiple outcomes, each distinctive and separately considered
2. an insistence upon the specification of these outcomes clearly and transparently, such that assessors can understand what is being assessed and what should be achieved
3. removal of the relationship between educational settings or learning programmes from assessment.

Outcomes-based assessment is often conducted through observation of performance on realistic tasks in a workplace setting. Portfolios are also common formats, in which evidence of performance on the assessment criteria are collated. Subject matter experts from the vocational sector are generally deemed the most suitable assessors. The assessment itself is often a list of criteria against which the assessor judges the learner's performance. Quality assurance is systematized by ensuring that the necessary procedures have been followed and that evidence has been logged appropriately. Verification exercises might include inter-rater reliability checks, but they emphasize record-keeping and processes to a larger extent since in this paradigm the judges are to be trusted to ascertain who is fit to practise on the basis of observed performances. Communities of practice, in which assessors come to understand through group interaction how to apply the criteria, are important theoretically to the outcomes-based paradigm (e.g. Lave and Wenger, 1991; Klenowski and Wyatt-Smith, 2014).

Outcomes are generally pass or fail categorizations rather than interval scales. After all, from this perspective the surgeon is either fit to practise or she is not. Given the purpose of the assessments, a high pass mark is often set to indicate that the learner has mastered the subject matter.

Unlike in a psychometrics tradition, there may be no scores, so there is no assumption of a normal distribution. The standard setting method is criterion-referencing, in which criteria are formulated that describe what the learner must know and be able to do. These are then applied to the observations or judgements of performances in portfolios or other tasks. Predictive validity of job performance is important in this paradigm, since the purpose of the assessment is to ensure competency to practise. A curriculum is usually specified by the occupational sector for which the assessment is designed. Chapter 15 discusses the implementation of outcomes-based assessment reforms in South Africa and New Zealand.

The main problem with this approach is that statistics play no role in a purely outcomes-based paradigm. This can cause havoc with an education system in which there are expectations of general stability from year to year in the cohorts taking national examinations. Concerns have also been raised about the impact of the outcomes-based approach upon learning, with some authors arguing that assessment comes to replace learning in some programmes. Torrance (2007) termed this 'assessment *as* learning' (p. 281). Equally, the lack of inter-rater reliability of standards judgements has been problematical, as this approach depends upon the experts who make the judgements being able to do so consistently (Wolf, 2002). With its atomistic approach to assessment, production of high quality rubrics that can support these judgements is problematical.

Curriculum-based assessment paradigm

From the field of education itself comes our final paradigm. In curriculum-based assessment, there is a lengthy tradition of school- and university-based examinations; in China the Imperial Examinations mentioned above have been around for centuries. Curriculum-based assessment has often been contrasted with psychometrics in the literature and simply termed 'examinations' or 'assessment' by authors such as Desmond Nuttall (1987) or Caroline Gipps (1994). Arguments for a curriculum-based assessment paradigm were closely connected with views about learning and the interaction between assessment and learning. Gipps (1994) indicated that assessment differed from psychometrics because it

- did not see learning as a fixed property of the individual, but as something malleable
- was criterion- rather than norm-referenced
- focused more upon validity in assessment design (whereas much of psychometrics perhaps unnecessarily erred on the side of reliability)

- relied upon formats that assess higher order thinking in depth
- was designed to produce the best performances from individuals with clearly presented, relevant, concrete tasks that were not overly anxiety-provoking.

Some of the above list is a reaction to examinations as well as psychometrics. Assessment for learning principles arose from the Assessment Reform Group in London, whose work built upon that of Nuttall and whose members included Gipps. Other classroom assessment movements are also consistent with the curriculum-based paradigm (e.g. Shephard, 2001).

The curriculum is defined by educational experts such as teachers and is usually disciplinary in nature. Rather than attempting to assess an underlying trait, the attribute of interest is performance on assessments, which is assumed to be caused by the knowledge and skills of the candidate gained through studying the curriculum. Outcomes are sometimes scores, but usually grades. Typical formats are written examinations, which may include extended answers or constructed response questions. The main purpose of the assessment is to give assurance that the individual has demonstrated that she has acquired enough knowledge and skill to progress to the next stage of education.

With the emphasis being upon the curriculum as a statement of learning goals, it is therefore important that the assessment aligns with the curriculum. Further, the assessment itself is essentially a more detailed indication of what students should learn. Therefore, transparency of assessments is very important to this tradition: past papers are often published so that students can see what they need to know and be able to do in the examination. Construct validity in relation to the curriculum is prioritized; a test would be seen as unfair if it assessed matters that went beyond the curriculum or did not properly represent the full range of the curriculum.

Because no score distributional assumptions are germane to this paradigm, tests are not redesigned if the scores are not normally distributed. Inter-rater reliability is emphasized because of the social function of the examinations and the need for fairness across schools. Selective functions of curriculum-based assessments are central to this approach. As such, standard setting has traditionally been cohort-referenced (in which a certain proportion of test-takers are awarded each grade). The curriculum, examinations and standards are set usually by involving subject matter expert educators, though policymakers are also often heavily involved.

For those coming from a psychometrics perspective, the curriculum-based approach can appear as a watered-down, less theoretical version. Some of the assumptions underlying the curriculum-based assessment approach are the same across the statistical models that underpin psychometrics and classical test theory; the latter of which could be said to be more associated with curriculum-based approaches (Baird *et al*., 2017). In both the psychometrics and curriculum-based paradigms, there is an underlying assumption that the scores for the questions will correlate with each other; that people who do well on one question will tend to do well on others. This has been taken as a formal indication that the assessment overall is addressing the same thing, rather than being a meaningless amalgam of unrelated factors. However, psychometric approaches are more stringent about the need for high correlations between items (internal reliability), and the curriculum-based approach instead prioritizes coverage of the syllabus. Therefore, interpretations of correlations between items, or internal reliability, differ between these paradigms. Recently, Maul (2013) has reconceptualized the underlying construct in curriculum-based assessment as composite variables. To illustrate, when assessing English, in a curriculum-based approach, low correlations between speaking, listening, writing and reading skills would not necessarily be deemed problematical so long as they addressed the knowledge and skills set out in the curriculum. These are matters of extent and emphasis. To reject issues of inter-item correlation entirely would be an extreme position.

Although this book is intentionally neutral about the supremacy of any one paradigm, we do not believe that assessment itself or the choice of paradigm is neutral. As Moss *et al*. (2005) stated:

> Different methods and theories have implications for the ways in which concepts such as learning or educational reform or fairness are formulated, studied and promoted through practical activity. Perhaps more profoundly and subtly, these methods and theories affect the ways human beings are represented and, ultimately the ways they come to understand themselves and others (Moss *et al*., 2005: 70).

Let us turn to a historical study that began nearly 90 years ago and had a similar structure to this book's project: the involvement of representatives from a range of countries, brought together to investigate educational assessment methods.

The International Examinations Inquiry

Between 1931 and 1938, the International Examinations Inquiry was conducted, aiming to improve examining across countries (Lawn, 2008). The participants of the International Examinations Inquiry were hand-picked as the leading researchers in the field. Representatives from England, France, Germany, Scotland, Switzerland, the US, Finland, Norway and Sweden attended three international meetings. Membership of the project included Philip Hartog and E.C. Rhodes, Charles Spearman, (the now notorious) Cyril Burt, Godfrey Thomson, Edward Thorndike, Isaac Kandel, Nils Lundquist and Jean Piaget. This 1930s project was funded by the Carnegie Corporation; its aim was to advance the science of educational assessment. Despite best efforts, it did not proceed as planned and instead of advancing scientific examining knowledge cumulatively, it ended in disagreement and disarray, with a plethora of approaches to educational assessment being outlined. Fundamentally different purposes and principles abounded, making it impossible for the Inquiry to make progress, and the project was abandoned. Lawn (2008: 23) argued that the Inquiry had an Americanizing hegemony as its guiding principle, dominating the intellectual traditions of examining in other nations. The American approach was, and still is, a psychometric tradition. Notwithstanding, other countries have had colonial influences on other countries' education systems. We return to this issue in Chapter 16.

The American hegemony incorporated the idea of psychometrics as a scientific advance on other practices. However, the German delegation at the International Examinations Inquiry was interested in the idea of education as individualistic self-cultivation (*Bildung*), which is somewhat at odds with psychometric traditions. In England, the power of the Oxford and Cambridge examinations might have meant that the English delegation found it difficult to change traditions, though the Inquiry's research on lack of consistency between examiners' marking did a lot to undermine public confidence in the system (Hartog and Rhodes, 1936). That legacy is still apparent in England. The Swiss delegation focused on the effects of national examinations on teachers and classroom practices; again, not questions that naturally arise through a psychometrics lens. For the French delegation, there was a split in which some participants saw that the testing and psychometrics tradition could narrow the curriculum and thereby reduce the *culture générale* that they were interested in examining through traditional methods. Work by the Norwegian delegation supported

the validity of the national examinations and therefore did not move the country in the direction of psychometric testing.

Lawn (2008) concluded that there were deep effects of the International Examinations Inquiry upon educational research and policy in many of the countries from which a delegation participated, though they may not have been the effects first intended by the project leaders or sponsors. By the end of the Inquiry, examining largely remained a national phenomenon that was culturally bound rather than international and objective. In particular, the modern science of examining, as the US delegates saw it, was not uniformly adopted.

The Inquiry took place at a time when the field of psychology was still establishing itself with scientifically credible methods. Assessment issues were central to those debates as the move from the study of unobservable, subjective, phenomenal experience to objective, observable and replicable measurements of people's behaviour was key. Not everyone was convinced that studying people's behaviour could be done in the same way as measurements could be made in the hard sciences such as physics. What were the units of measurement? Were they stable across conditions and over time? Could psychological data only ever be qualitative? The Ferguson Committee (Ferguson *et al.*, 1940), set up by the British Association for the Advancement of Science, provoked Stevens' 1946 argument (Stevens, 1970), that there are different scales of measurement (categorical, ordinal, interval and ratio), all of which were useful. These scales and the debates surrounding them are still contentious today in educational assessment, so let us explain them. They are hierarchical in nature, with the properties of the preceding scale being subsumed into that of the next. For example, every ratio scale also has the properties of interval scales, ordinal scales and categorical scales.

Categorical scales permit only the classification of things. Examples include colour, occupations, gender or nationality. Some types of data permit more than categorization, since there is an ordering to the categories; these are ordinal scales. Examples include occupational indexes in which jobs have been ranked in terms of salary, responses to rating scales and socio-economic status. Ordinal scales do not have equal measurement units; only rank ordering of the units is a feature. In interval scales, the units have equal intervals as well as being rank-ordered. An example of an interval scale would be temperature measured in Celsius. Consistency of the intervals between Celsius degrees is scientifically meaningful in terms of the states of water. One feature of such scales is that they do not have an absolute zero, which renders certain calculations involving multiplication or

division meaningless. It makes no sense to any external referent to say that ten degrees is half the temperature of twenty degrees. A ratio scale is one in which the intervals are meaningful and there is an absolute zero. The euro is an example of a ratio scale; having no money is very meaningful, as is having double the amount of money you started off with. In educational assessment terms, a pass grade (with no others available) would be categorical, letter grades would be ordinal and scores are interval. Arguments about the use of scores as interval data persist; some argue that they are at best ordinal data. In practice, of course, we often treat examination results as though they are interval data and are quite happy to construct mean point scores.

Psychometrics has attempted to construct interval scales from psychological and educational data (specifically item response theory techniques). Interval scales are very powerful because they can build upon the voluminous advances that have been made in statistical methods. Modern statistical techniques are also available for all forms of Stevens' scales, though there are technical requirements associated with each and some are less straightforward than others.

Our Standard Setting Project was somewhat more diverse than the 1930s International Examinations Inquiry US–Europe project. Participants were from Australia (Victoria, Queensland), Chile, Cyprus, England, France, Georgia, Hong Kong, Ireland, Italy, Norway, Scotland, Singapore, South Africa, Sweden, the US and Wales. Given the International Examinations Inquiry outcomes, the fact that different practices coexist and the debates in the research literature, we anticipated significant disagreements over what counts as sound practices. That is what we encountered.

The aim of our Standard Setting Project was not the same as the International Examinations Inquiry, as we did not seek to promulgate best practices, but to depict and compare approaches for the important national examinations that are held in these countries. We were open to the notion that different views on examination philosophy might exist. Understanding the meaning of the standard setting practices from policy documents alone is problematical precisely because the practices are embedded within wider cultures and structures. Therefore, we sought to find out why particular practices were more acceptable in their context than others, what the meaning of examination standards were in their context and whether the approaches could be classified. Certainly, one important finding is that there are many national examination systems that operate outside of the psychometrics tradition. School leaving examinations in England and Scotland were known examples (Baird and Gray, 2016), but there turned out to be many more.

One perspective arising in the Standard Setting Project, but detectable more broadly, is that a great deal of educational assessment practice, including in classrooms and examinations all over the world, is inferior in nature and would benefit from modernization, using psychometric techniques. This is a common thread from the International Examinations Inquiry. Ways in which psychometrics could assist these practices can be readily envisaged, but they may come at a cost and they may be too much of a distraction from the central purposes of the assessments (Baird and Black, 2013). Indeed, a challenge for those who take the scientific-psychometrics-superiority view is to explain why, if this really is so much better an approach, has it not simply been adopted in a blanket fashion. Is it lack of expertise, cost or woolly thinking? An alternative argument, posited here, is that there are three different ways of construing educational assessment. Each has its advantages and limitations, and the social process of moving from one paradigm to another would, in itself, be political and complex, as discussed in Chapter 15.

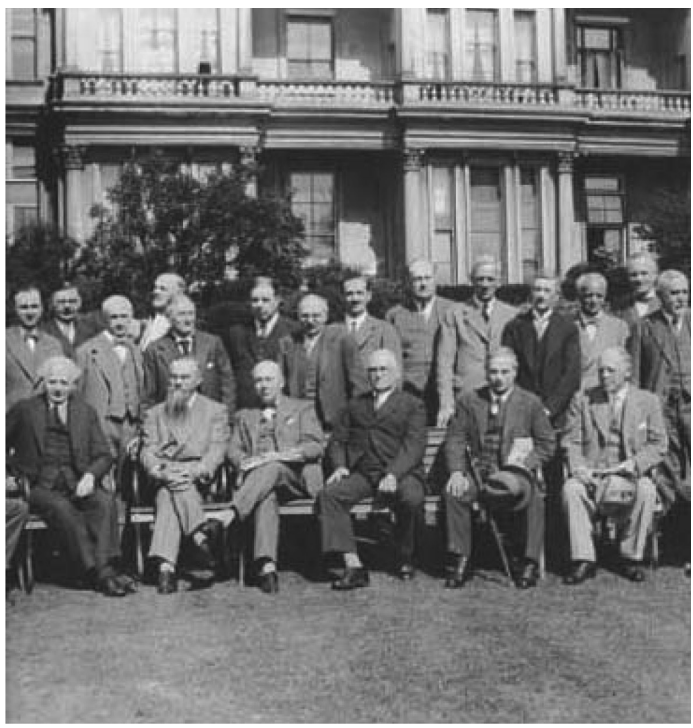


Figure 1.1: Some International Examinations Inquiry attendees (1930s)

Figure 1.2: Brasenose Standard Setting Project Symposium attendees (2017)

Background to the Standard Setting Project

This book is the culmination of a collaborative project on international standard setting between the Oxford University Centre for Educational Assessment (Professor Jo-Anne Baird), the Assessment and Qualifications Alliance (Dr Lena Gray), UCL Institute of Education (Dr Tina Isaacs) and Ofqual (Dennis Opposs). The overall project included contributions from 12 jurisdictions across the developed and developing world to a symposium held in Oxford in March 2017. The book explores the trenchant themes emanating from those contributions and highlights case studies from nine of them, chosen to illustrate different systems that are in use. The project's full title is: Setting and Maintaining Standards in National Examinations. We normally refer to it in this book simply as the Standard Setting Project.

The research aims for the project were to investigate, document, analyse and evaluate four key aspects of national standard setting systems:

- how standards are defined in national curriculum-related examination systems, whether they be school leaving or university entrance
- how those definitions are enacted in terms of processes and evidence used
- issues for the system and responses to those issues
- the commonalities and diversity of definitions of, processes for and challenges to standards.

Assessing the achievement of curriculum standards is powerfully enacted through processes of standard setting and maintaining within curriculum-related examinations. Many jurisdictions use curriculum-related examinations to select learners for higher education, work and other study options. Some

jurisdictions also use these examinations as tools to measure the performance of their schools. As such, these examinations shape the landscape of senior school education, defining the quality of education system for learners and for society.

The focus of the Standard Setting Project, and thus of this book, is national, school leaving or university entrance, curriculum-related examination systems from a range of jurisdictions around the world. It aims to describe the processes used to set or to maintain (or to link over time) standards in these examinations and to explore the concepts relating to standards behind them. These examinations are particularly important for the young people that take them. Each year around the world, tens of millions of young people take these types of examinations. For most of them, the result they receive from their examination will be an extremely important determinant of where they progress to in terms of education or employment.

Given the high stakes of these examinations, it is surprising that the ways examination standards are conceptualized and operationalized differently across jurisdictions have not been given more attention. Very little has been written that documents and conceptualizes the meaning of examination standards in high stakes national examinations. In England, although the meaning of examination standards has been much debated in the literature (Baird, 2007; Baird *et al.*, 2000; Cresswell, 1996; Christie and Forrest, 1981; Coe, 1999, 2007, 2010; Newton, 1997a, 1997b, 2003, 2005, 2010), it is often noted that stakeholders discuss examination standards using contradictory definitions without realizing they are doing so. Thus, more clarity is needed in the field; one purpose of this book is to contribute to that.

While most national examination systems use both statistics and examiner judgement in their standard setting processes, a lack of transparency often characterizes how various sources of information are used in the decision making. This is an interesting area because although globalization has begun to impinge on examination systems, public examination standards are still largely a bastion of the local. Educational cultures differ across jurisdictions, affecting assessment structures and processes in distinctive ways. The meaning of 'standards' differs between jurisdictions, and the stated value positions and processes relating to examination standards differ markedly.

How policy and politics affect standards across different jurisdictions has not been well articulated. Further, there is a tenuous relationship between standard setting theory and the manner in which jurisdictions operationalize

standards setting. The under-articulation of the rationales for current examination practices (including standard setting practices) means that they are vulnerable to changes that could well be detrimental to the character of their education systems. This book should help inform future developments by making clear to researchers, policymakers and practitioners interested in assessment the definitions of examination standards, describing how they are operationalized and explaining what impacts the definitions have upon how standards are interpreted in the wider community and in the media.

In this book we aim to examine critically policy positions and processes for setting standards in a range of jurisdictions. The project aims to illuminate similarities and differences in conceptual bases and operational approaches to standards through both thematic and case study chapters. It challenges current theory on standards, and may lead to changes in how national organizations approach standard setting and maintaining. For the first time to our knowledge, the research on examination standards definitions reaches beyond a single country or a comparison of a small number of countries.

As well as practices differing between jurisdictions, so too does the use of language. Sometimes the same concept has different names in different places. So, for example, in most of the jurisdictions in the Standard Setting Project, the written assessments that students take are called 'examinations'. However, in some jurisdictions (Chile, South Korea and Sweden of those within the project), when writing or speaking in English the same form of assessment is referred to as a 'test'. In most countries, the name for the lowest possible mark in an examination that a student must achieve in order to gain a particular grade is a 'cut score'. In the context of most examinations in England, the same concept is called a 'grade boundary'. We have permitted authors to use the words with which they are most familiar when writing contributions for this book, and we are not providing a glossary of terms. That does require readers to be aware that the use of different words in different chapters does not always indicate the use of different concepts.

Chapters of the book

Part One. Researching national examination standards

In this first chapter, we have explained why the members of the project board thought that the aims of the Standard Setting Project were compelling enough to deserve the resources required. We provided some relevant background to the principles that lie behind assessment practices. In a key section, we then identified three ways of construing educational assessment;

three paradigms. Within each of these paradigms, we introduced distinctive ways of thinking about educational assessment. These address and suggest different questions and ways of interpreting the evidence that is collected about them. Each has its advantages and limitations. In our view, the social process of moving from one paradigm to another would be political and complex.

The next three chapters of the book are thematic, each having a different focus but all drawing on the evidence generated during the project. Although each is distinct, there are important links between them. Chapter 2 describes the methodology of the project, so that the reader can come to a view about the quality of the work. In particular, for a project of this kind, authenticity and positionality issues needed to be addressed. In addition to the methodology described in Chapter 2, we consulted assessment experts on the research design, as indicated in the acknowledgements. Given our stance, it can be assumed that chapter authors have their own views on the matters raised in this book. Further, authors' positions are not necessarily the policy position of their employing organization.

In Chapter 3 we look at how to mitigate the risks when researching standards using insiders as a key source of evidence. We realize that an important contribution to the field could be made by codifying the political and organizational barriers to such work and delineating a range of ways in which individuals and organizations could overcome them to advance their national examination technologies and policies. Arising from this project, guidelines have been produced to enable examination board researchers to be more transparent about the procedures that they use and the challenges that they face. These can be found in Appendix B.

After clarifying the term 'standard setting', different methods of setting standards are classified and discussed in Chapter 4. For the first time, we relate the practice of combining different sources of evidence, or using both quantitative and qualitative data, that is common in educational assessment, to mixed methods methodology used in the social sciences. Finally, we investigate which methods each of 12 jurisdictions uses to set standards in its national, school leaving, or university entrance examinations.

Part Two. Case study chapters

Each of the nine chapters in Part Two focuses on a particular case study jurisdiction that formed part of the Standard Setting Project. We look at a key examination system in each of Chile, England, France, Georgia, Ireland, Queensland (Australia), South Africa, Sweden and the US. Each chapter follows a similar structure. After background about the jurisdiction

itself, we provide an outline of the national examination that forms the focus of the study. We describe assessment arrangements as well as the processes used to set standards. Finally, the chapters discuss some political and public controversies and debates about the examination. Since detailed information about the examination and its standard setting process is usually hard to find in other literature, we hope readers will find these chapters to be a valuable resource. Given that the issues of positionality and insider researchers were germane to the project, we also include two commentaries for each case study chapter, written by established assessment researchers who understand the context of the examinations. Commentary authors were free to address any relevant points that they considered would add to the discussion about standard setting for the examination in question.

Part Three. Differing measures and meanings

Chapter 14 investigates the different meanings of 'examination standards' that have previously been published in the literature. We rationalize the literature by using an ecological model to show that the definitions are associated with different levels of education and examining systems. Here, we show that criterion-, cohort-, construct-, attainment- and standards-referencing were all used in the examination systems participating in the Standard Setting Project.

Chapter 15 explores how standard setting processes fit and work in the wider political, social and cultural context. First, we analyse how accepted standards setting practices become enshrined through culture and context. Drawing largely on the evidence provided by project participants, we describe some examples of the ways different countries use national examinations in practice and present a framework to explain why fundamental change to national examination systems is so rare. We conclude that changes in standards setting are usually accommodations to existing models rather than paradigm shifts.

Chapter 16 then draws some conclusions about what all of this means for the state of the field in terms of theory, practice and policy. It highlights trends in issues that relate to standard setting such as trust (or lack of it) in the examinations systems, the role of social justice in standard setting and the role of teachers setting standards through teacher-based assessment. It then summarizes the contributions that the book has made to the standards setting literature both within the thematic and case study chapters. Finally, the conclusion addresses limitations to the research and areas for future research.

References

Andrich, D. (2004) 'Controversy and the Rasch model: A characteristic of incompatible paradigms?' *Medical Care*, 42, 7–16. Reprinted in E.V. Smith and R.M. Smith (eds), *Introduction to Rasch Measurement: Theory, Models and Application*. Maple Grove, MN: JAM Press. Chapter 7, 143–66.

Baird, J. (2007) 'Alternative conceptions of comparability'. In Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority, 124–56. Online. https://goo.gl/8SvTBo (accessed 7 June 2018).

Baird, J. and Black, P. (2013) 'Test theories, educational priorities and reliability of public examinations in England'. *Research Papers in Education*, 28 (1), 5–21.

Baird, J. and Gray, L. (2016) 'The meaning of curriculum-related examination standards in Scotland and England: A home–international comparison'. *Oxford Review of Education*, 42 (3), 266–84.

Baird, J., Johnson, S., Hopfenbeck, T.N., Isaacs, T., Sprague, T., Stobart, G. and Yu, G. (2016) 'On the supranational spell of PISA in policy'. *Educational Research*, 58 (2), 121–38.

Baird, J., Andrich, D., Hopfenbeck, T.N., Stobart, G. (2017) 'Assessment and learning: Fields apart?'. *Assessment in Education: Principles, policy & practice*, 24 (3), 317–50.

Christie, T. and Forrest, G.M. (1981) *Defining Public Examination Standards* (Schools Council Research Studies). Basingstoke: Macmillan Education.

Coe, R. (1999) 'Changes in examination grades over time: Is the same worth less?'. Paper presented at the British Educational Research Association Annual Conference, University of Sussex, 2–5 September. Online. https://goo.gl/RhCtmt (accessed 7 June 2018).

Coe, R. (2007) 'Common examinee methods'. In Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority, 331–67. Online. https://goo.gl/VvyVHG (accessed 7 June 2018).

Coe, R. (2010) 'Understanding comparability of examination standards'. *Research Papers in Education*, 25 (3), 271–84.

Cresswell, M.J. (1996) 'Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches'. In Goldstein, H. and Lewis, T. (eds) *Assessment: Problems, developments and statistical issues*. Chichester: John Wiley, 57–84.

Ferguson, A., Myers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W., Campbell, N.R., Craik, K.J.W., Drever, J., Guild, J., Houstoun, R.A., Irwin, J.O., Kaye, G.W.C., Philpott, S.J.F., Richardson, L.F., Shaxby, J.H., Smith, T., Thouless, R.H. and Tucker, W.S. (1940) 'Quantitative estimates of sensory events: Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events'. *Advancement of Science*, 1, 331–49.

Gipps, C.V. (1994) *Beyond Testing: Towards a theory of educational assessment*. London: Falmer Press.

Goldstein, H. (1979) 'Consequences of using the Rasch model for educational assessment'. *British Educational Research Journal*, 5 (2), 211–20.

Goldstein, H. (2012) 'Francis Galton, measurement, psychometrics and social progress'. *Assessment in Education: Principles, Policy & Practice*, 19 (2), 147–58.

Goldstein, H. and Wood, R. (1989) 'Five decades of item response modelling'. *British Journal of Mathematical and Statistical Psychology,* 42 (2), 139–67.

Grek, S. (2009) 'Governing by numbers: The PISA "effect" in Europe'. *Journal of Education Policy*, 24 (1), 23–37.

Hartog, P. and Rhodes, E.C. (1936) *The Marks of Examiners*. London: Macmillan.

Jessup, G. (1991) *Outcomes: NVQs and the emerging model of education and training*. London: Falmer Press.

Klenowski, V. and Wyatt-Smith, C. (2014) *Assessment for Education: Standards, judgement and moderation*. London: SAGE Publications.

Kuhn, T.S. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lave, J. and Wenger, E. (1991) *Situated Learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lawn, M. (ed.) (2008) *An Atlantic Crossing? The work of the International Examination Inquiry, its researchers, methods and influence* (Comparative Histories of Education). Oxford: Symposium Books.

Maul, A. (2013) 'Method effects and the meaning of measurement'. *Frontiers in Psychology*, 4, 1–13.

Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) (2007) *Techniques for Monitoring the Comparability of Examinations*. London: Qualifications and Curriculum Authority. Online. https://goo.gl/6kqoXe (accessed 7 June 2018).

Newton, P. (1997a) 'Examining standards over time'. *Research Papers in Education,* 12 (3), 227–47.

Newton, P.E. (1997b) 'Measuring comparability of standards between subjects: Why our statistical techniques do not make the grade'. *British Educational Research Journal*, 23 (4), 433–49.

Newton, P.E. (2003) 'Contrasting definitions of comparability'. Paper presented at the QCA Standards and Comparability Seminar, Milton Keynes, April.

Newton, P.E. (2005) 'Examination standards and the limits of linking'. *Assessment in Education: Principles, Policy & Practice*, 12 (2), 105–23.

Newton, P.E. (2010) 'Contrasting conceptions of comparability'. *Research Papers in Education*, 25 (3), 285–92.

Neumann, W. (1979) 'Educational responses to the concern for proficiency'. In Grant, G., Elbow, P., Ewens, T., Gamson, Z., Kohli, W., Neumann, W., Olesen, V. and Riesman, D. *On Competence: A critical analysis of competence-based reforms in higher education*. San Francisco: Jossey-Bass, 6–94.

Nuttall, D.L. (1987) 'The validity of assessments'. *European Journal of Psychology of Education*, 2 (2), 109–18.

Porter, T.M. (1986) *The Rise of Statistical Thinking, 1820–1900*. Princeton: Princeton University Press.

Roberts, J.A.G. (2006) *A History of China*. 2nd ed. Basingstoke: Palgrave Macmillan.

Shepard, L.A. (2000) 'The role of assessment in a learning culture'. *Educational Researcher*, 29 (7), 4–14.

Spady, W.G. (1977) 'Competency based education: A bandwagon in search of a definition'. *Educational Researcher*, 6 (1), 9–14.

Stevens, S.S. (1970) 'On the theory of scales of measurement'. In Haber, A., Runyon, R.P. and Badia, P. (eds) *Readings in Statistics*. Reading, MA: Addison-Wesley.

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), 281–94.

Tuxworth, E. (1989) 'Competency based education and training: Background and origin'. In Burke, J. (ed.) *Competency Based Education and Training*. Lewes: Falmer Press, 10–25.

Tyler, R.W. (1949) *Basic Principles of Curriculum and Instruction*. Chicago: University of Chicago Press.

Wolf, A. (1995) 'Competence-based assessment'. In Raven, J. and Stephenson, J. (eds) *Competence in the Learning Society*. New York: Peter Lang, 453–66.

Wolf, A. (2002) *Does Education Matter? Myths about education and economic growth*. London: Penguin Books.

Chapter 2

Researching national examination standards: Our methods

Lena Gray

One of the key issues to consider as we embarked on the Standard Setting Project was what research techniques could be used to ensure that we had the fullest, most accurate picture of any standard setting system. We recognized, as insiders to the industry, that the collation of formal policy statements on standards would not be sufficient, as practice can differ from stated policy. Authenticity of the research was clearly important; however, as this chapter indicates, there are many aspects and layers to authenticity.

To investigate how standards are set in a range of countries, case study methodology was necessary. We decided to adopt a multiple-case embedded model (Yin, 2014: 50); each case has its own contextual conditions, but with multiple units of analysis within each one. We selected this methodology because, in Yin's words:

> Case study research would be the preferred method, compared to others, in situations when (1) the main research questions are 'how' or 'why' questions; (2) a researcher has little or no control over behavioural events; and (3) the focus of study is a contemporary (as opposed to entirely historical) phenomenon (Yin, 2014: 2).

Our use of a multiple-case approach was intended to 'shed empirical light about some theoretical concepts or principles' (Yin, 2014: 40) by comparing cases that mirror and confirm existing documented definitions of standards with contrasting cases, and so provide a challenge to those documented definitions and allow us to move thinking forward. Our chosen cases, then, were selected to cover distinctive approaches to standard setting, geographical spread, cultural distinctiveness, different assessment formats and use of differential cut scores from the same examination.

Having selected a case study method and a multiple-case design for our study, our next methodological challenge was to try to establish what

our sources of data would be. Analysis of documents and archival records form part of each case study, embedding different units of analysis within each. A pilot study on standard setting in Scotland and England indicated that documentary and archive evidence is not enough on its own, as publicly documented positions on standards can be too brief, contradictory, outdated, or may not reflect practice in other ways (Baird and Gray, 2016). We needed therefore to use more than one source of evidence; the sources we selected are summarized in the diagram below.

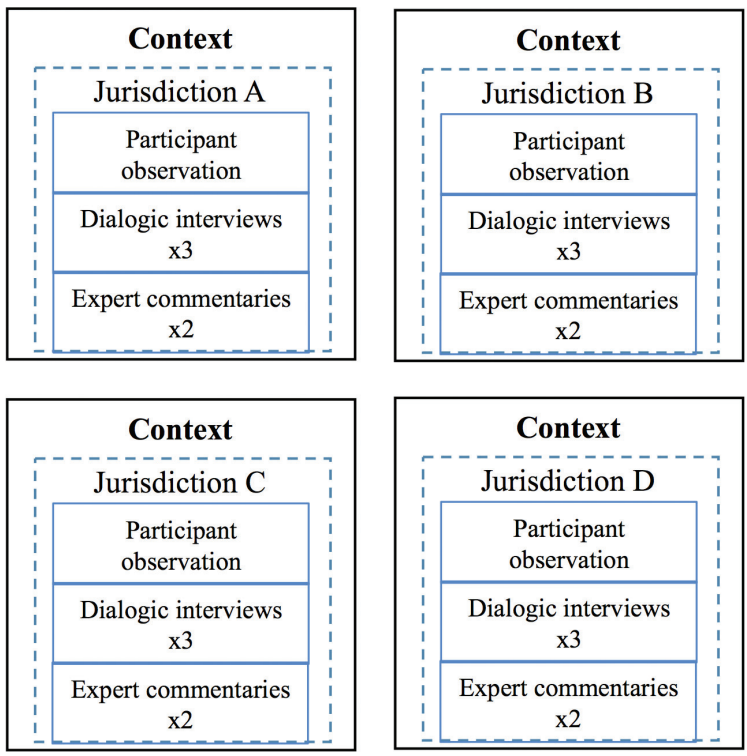


Figure 2.1: Multiple case embedded model adapted from Yin, 2014

The methodological literature on qualitative research has, in the past, suggested that gathering data from additional sources would help to validate the research findings by allowing triangulation of the data (Creswell, 1997: 202, summed up the literature to the time of writing); later methodological texts, though, suggest that such advice stemmed from reactions to positivistic critiques of qualitative research (Morse, 2018) and fail to acknowledge the premise that it is not possible to arrive at a 'context-free "truth"' (Silverman, 2011: 443). We situate our research in that more recent theoretical frame, and thus our use of multiple data sources is not 'for reliability and to ensure

replication' but, instead, it 'ensures comprehensiveness of the topic and domain' (Morse, 2018: 804). As such, the authenticity of our findings is assured not only by pre-research design and planning, and post-research validation checks, but, more importantly, by being 'cohesively embedded in the method used, as they move the analysis forward' (Morse, 2018: 799). Within this more recent methodological literature, terminology related to quality is disputed and shifting, although arguably it would be fair to say that the most recent proposed frameworks exhibit a shift away from quasi-technical terms like reliability and validity and instead appear to be aiming to use broader language, adopting simpler words such as 'quality' (Tracy, 2010) and 'rigor' [*sic*] (Morse, 2018). We have chosen to use the terms 'authenticity' and 'trustworthiness' to connote the basis on which we invite readers to judge the quality of our work. We have not provided separate definitions of these terms, but instead we explore their meaning in the context of this research, both throughout this chapter and Chapter 3.

Through a multi-layered research design, we sought to ensure the authenticity and trustworthiness of the data that we gathered, using the most pertinent of the approaches identified in the methodological literature (e.g. Creswell and Miller, 2000; Shenton, 2004; Morse, 2018). Shenton (2004) recommended a wide range of strategies for ensuring trustworthiness in qualitative research projects. We judged several of these to be relevant to our project and employed them throughout the research. These strategies included: ensuring that we as researchers had credible backgrounds, qualifications and experience; taking early steps to establish rapport with our project participants; using a variety of questioning techniques to facilitate openness; drawing on a wide range of data sources; providing opportunities for peer scrutiny of data; using member checks; and, throughout the research project, taking time to reflect on our own researcher roles and performance (Shenton, 2004: 65–8). The rest of this chapter expands on how we implemented these strategies.

Our first decision as to how to ensure authenticity and trustworthiness was in relation to our data sources. As we have noted, we decided to gather data from a range of sources, not in order to triangulate and arrive at an empirical truth, but in order to add breadth and depth to our research. We considered using interviews and direct observation as our main data-collection techniques, but given the international nature of the project, these approaches would have required significant resources. Additionally, these methods were likely to suffer from a lack of depth of understanding of the educational and assessment environment on the part of the researchers, not to mention language skills. Participant observation was therefore adopted

as the main data source, involving participants who already worked within the systems under investigation. This method provided for immediacy of data, and gave us access to evidence that may not otherwise have been available. As Yin pointed out:

> Participant-observation provides certain unusual opportunities for collecting case study data, but it also involves major challenges. The most distinctive opportunity is related to your ability to gain access to events or groups that are otherwise inaccessible to a study (Yin, 2014: 116).

Using participant observers who were already part of the jurisdiction under investigation meant that we needed contributors who knew the system well, and who were able to discuss publicly and document issues that could be controversial in their own context. We mainly selected senior personnel in exam boards to write the case study chapters. This did, however, threaten to limit the findings in a number of ways; in particular, the case studies may cover the policy intentions, rather than the lived reality. While practitioners who set the standards have accurate and up-to-date knowledge of policy and technical issues, they are also constrained by commercial interests and national politics in fully disclosing this knowledge. Even in setting out the official position, there may be variations in how full a description can be provided in each case study chapter. Some systems have complete policy statements, and perhaps public documents and archives on standard setting processes, but in other systems, this information may not be in the public domain, and an exam board employee may not be in a position to release it into the public domain. Using participant observers who are already part of the jurisdiction under investigation guaranteed that contributors had good knowledge of the system, and gave us the best chance of ensuring that we had the fullest picture of standard setting systems. However, it raised the issue of how to ensure that the picture presented was unbiased. In other words, how could we ensure that the project's findings would be viewed as trustworthy? We return to this issue below.

A further concern was that the limitations of participant observation may have been compounded by the issues facing exam boards and assessment bodies. As Baird and Gray (2016) suggested, 'Examination boards have a tricky, political task in managing public and stakeholder perceptions of examination standards' (Baird and Gray, 2016: 2) and therefore exam board personnel may not be able to discuss public critiques of their system in a full and open way. Whatever our research methodology, confidentiality could not be assured to participants: the participants (or in this case, their

organizations) have a high profile and are identifiable. Chapter 3 and Appendix B deal with these issues more fully. It was central to our research design that the organizations should be identifiable and that the research should include and analyse issues that are subject to public debate. We know from comparative work in education that the policy and educational reform landscape is constantly moving in many jurisdictions around the world (OECD, 2015). Such policy shifts may create space that allows discussion of the strengths and weaknesses of different systems, processes and concepts of standards; on the other hand, policy reform can lead to 'resistance from policy-makers to listening to the concerns raised by education and assessment professionals' (Baird and Coxell, 2009: 114), stifling debate and making it difficult for professionals like exam board employees to articulate their knowledge publicly.

Despite these tensions, there was little doubt that senior exam board personnel are the people who have the knowledge to provide a full and accurate description of their systems, their underlying principles and how they work in practice. Initially, we asked the senior exam board participants to compose papers for the project's invited symposium (described below). The papers from the different jurisdictions were required to follow a template, which we had prepared with the aim of making them as similar in structure as possible. While authors were asked to base their papers on this template, some adapted it to better suit their own contexts. In the Georgia case study, for example, there is only a short section about the technicalities of standard setting; in the Queensland case study, there is more focus on future assessment reforms than on description of the current system. In both cases, we were more interested in what standard setting means in the context and what problems have been encountered than in a narrowly technical description of the current system. The project team worked with case study authors through dialogue and a process of co-creation, engaging with the authors at all stages of the writing process, providing feedback, open dialogue and challenge in preparation for, during and following the symposium.

As mentioned above, we knew that using interviews as a primary source of data would bring a number of challenges, including those associated with power asymmetries (Brinkmann, 2018: 588) and those arising from the researchers' status as 'insider-outsider' interviewers (Mercer, 2007). However, so as to give our project participants a more active and primary role as producers of their own accounts, we did decide to use interviews as a secondary source of data, accepting Brinkmann's (2018) characterization of interviews as 'humane, inter-subjective and responsive' (578). In particular,

we wanted to make use of the dialogic benefits of the semi-structured interview so that we, as researchers and project participants, could jointly engage in 'knowledge-producing' (ibid.: 579) through 'an interchange of views between two or more people on a topic of mutual interest' (Cohen *et al.*, 2017: 506). In deciding to include interviews in our research design, we recognized 'the centrality of human interaction for knowledge production' and the 'social situatedness of research data' (ibid.: 349). Thus, while acknowledging the possible limitations of interviews as a means of data collection (Wragg, 1994: 267; Denscombe, 2010: 190), we designed our study to include successive rounds of what have been called 'dialogic interviews':

> **Dialogic interviews** are true conversations in which researcher and participant together develop a more complex understanding of the topic. There is authentic give and take in these interviews – mutual sharing of perspectives and understandings – and 'talk time' is more balanced between researcher and participant (Rossmann and Rallis, 2003: 182, emphasis in original).

We conducted two dialogic telephone interviews with each of the authors, plus a face-to-face interview at the symposium. These interviews provided us with a further source of data that allowed us to challenge our own understanding and that of our project participants, as part of a collaborative process of knowledge production. We have drawn on the interview data in drafting the thematic chapters of this book.

The first interview was used in part for rapport building and to establish the project rationale and parameters. In this conversation, we provided information about the symposium, introduced the idea of critical friends as commentators to be involved after the symposium, discussed the formal consent required for participation in the project and ensured that the author considered possible consequences of their participation in the project for them personally, and for their organization within their jurisdiction. (We obtained ethical approval for the project using the ethical procedures of each of our organizations. The University of Oxford's Research Ethics Committee approved procedures and AQA's Research Code of Practice both require adherence to the British Educational Research Association (BERA) Ethical Guidelines for Educational Research (2011).) In one sense, the purpose of the first interview was to explain the planned research methods and to ensure informed consent on the part of the project participants; we also talked through the symposium paper template to investigate whether the template provided an appropriate framework to elicit the data we

sought. Importantly, this first interview was an opportunity to establish the credibility of our own experience and backgrounds, as part of the process of establishing rapport with our project participants, as Shenton (2004) advised. Each member of the project team was assigned three or four project participants with whom he or she would work throughout the project. Wherever possible, we matched ourselves with jurisdictions with which we had a degree of familiarity.

Another aim of the first interview was to begin the dialogue about standard setting methods and definitions. Most of our project participants worked for exam boards – not always as researchers, but in senior roles with operational responsibility. We recognized that they would not necessarily have been familiar with the academic literature on standard setting; several expressed some concern that they were not academics, and did not ordinarily analyse their own practice, especially not in a theoretical way. We wanted to enable participants to do this, so that by the time they attended the symposium, they would feel comfortable discussing theoretical aspects of their own and others' standard setting systems. During the first interviews, our attempt to familiarize participants with analysing their own practice took the form of an initial discussion about standard setting. This was based around the requirements outlined in the symposium paper template, which they had received prior to the interview (along with the paper by Baird and Gray, 2016 manuscript) to allow them to prepare for the interview discussion. Following the first interview, we sent each participant a short briefing paper that aimed to summarize the established research on standard setting.

We followed these initial interviews with a second, more in-depth telephone interview a month or two later. This semi-structured interview was organized around discussion of questions such as: Who is responsible for standard setting and maintaining in your context? How are standards set in your context? What standard setting techniques are used? Do you use norm-referencing, criterion-referencing, attainment-referencing or another method? What are the controversies in your context around examination standards and what do they tell us about standard setting and maintaining? In this discussion, we drew on the insights and definitions that had been outlined in the briefing paper, and asked participants to describe their own practice and reflect on it through the theoretical lenses provided by the academic research on standard setting. The principal aim of this round of interviews was to allow us to draw out and challenge the participants' understanding and accounts of their own systems, as a collaborative aid to their, and our, self-reflective analysis, since 'Credible data also come from

close collaboration with participants throughout the process of research' (Creswell and Miller, 2000: 128).

Our interviews were an important means of mitigating the risks of limitations in research data gained from the authors, allowing us to discuss that data in a process of dialogic knowledge production. In order to further strengthen the authenticity and trustworthiness of our research, we decided that the project would also encompass alternative perspectives:

> An investigator must seek those alternatives that most seriously challenge the assumptions of the case study. These perspectives may be found in alternative cultural views, different theories, variations among the stakeholders or decision-makers who are part of the case study, or some similar contrasts (Yin, 2014: 204).

As well as including alternative analyses of cases in the overarching chapters, which draw out key themes from across the case studies, we accessed a range of informants to provide alternative perspectives that might pose rival explanations of the phenomena described. We asked additional in-country experts to provide commentary and analysis on policies and processes relating to standard setting and maintaining within the jurisdiction. The experts were given the relevant case study chapter and asked to respond to it, including any insights or critiques that may be different from the chapter author's analysis. The commentaries provided a means to address the limitations of bias and possible insider researcher difficulties with disclosure. Although these commentaries were later shared with the chapter authors, authors were not given the chance to amend their text or to respond to the commentaries; however, they could raise issues of factual inaccuracy. In the event, this never happened. The commentaries follow the case study chapters to which they relate.

An important part of the research process was an invited symposium held at Brasenose College, Oxford, in March 2017, at which 46 colleagues from a wide range of jurisdictions presented work on what examination standards mean in their context. Delegates were carefully selected to represent researchers who had published important work on educational standards and those who were responsible for examination standard policy and practice for national examinations. We provided papers in advance from each of the 12 jurisdictions that form part of the Standard Setting Project, and each project participant (or, in two instances, their representative) presented to the invited audience of experts on a key theme associated with their case. In this way, the symposium provided an opportunity for peer scrutiny of participants' accounts and analyses of their own systems, and formed

another important means to secure trustworthiness in our research findings. It also contributed to one of the secondary aims of the project, which was to establish a knowledge community. Although there are established networks that are used by senior exam board colleagues, academics and policymakers, such as the International Association for Educational Assessment (IAEA) and the Association for Educational Assessment – Europe (AEA–Europe), those who have responsibility for implementing standard setting policies are not always represented in these networks. All of our participants expressed a wish to learn from other systems and valued the symposium as an opportunity to do so as a first step to creating a knowledge community.

The interviews conducted at the symposium provided further data about issues that had been raised in some of the plenary sessions, such as what definition of standards would provide a good description of the curriculum-related examination operating in their jurisdiction, and who has the power (either hard or soft) to define and set standards in their jurisdiction. Using an iterative interview design, with three rounds of interviews, allowed us to help participants to reflect on their own experience from a position that opened up fresh viewpoints on the processes they had known intimately; this was especially true of the interviews that took place after the extensive presentations and discussions at the symposium. After each of the three interviews, records were shared with participants in a process of 'member checking' (Creswell and Miller, 2000). This process allowed participants to confirm that the interview records captured their intentions and strengthened the authenticity of the findings by providing an opportunity for the project team to verify 'the investigator's emerging theories and inferences as these were formed during the dialogues' (Shenton, 2004: 68). An additional form of member checking took place when the chapters of this book reached a late, draft stage; case study commentaries were shared with relevant participants, and thematic chapters were shared with all project participants so that they could 'confirm the credibility of the information and narrative account' (Creswell and Miller, 2000: 127).

The interviews, and the presentations, papers and discussion that took place at the symposium indicated the wide range of practices in use and provided a major source of data for this book.

Insider research

It would be remiss of us, in discussing the methodology of the Standard Setting Project, not to explain how we dealt with insider research issues. Chapter 3 of this volume discusses broader, more theoretical insights into insider research that were gained during the course of the project. The

remainder of this chapter briefly clarifies how these issues affected project participants during the project; it draws on the records of the three rounds of interviews to illustrate some of the points made.

The project was preceded by a pilot project, in which Baird and Gray (2016) focused on a comparison of curriculum-related examination standards in Scotland and England. Methodologically, the initial project used critical evaluation of published policy documents and the authors' insider experiences of standard setting in Scotland and England (Sikes and Potts, 2008). Both authors had had professional responsibility for standards in an English examining board, and one of the authors was formerly responsible for standard setting policy at the Scottish Qualifications Authority (SQA). Their depictions of the standards policies in Scotland and England were constructed in part through member checking with senior exam board personnel responsible for standards: the authors presented their interpretations of the stated policies and discussed these with the senior practitioner/policymakers as part of a collaborative contribution to the field (Creswell and Miller, 2000: 126). The authors noted their own status as insider–outsiders and the effects of this on the research.

In initiating the international project, such insider–outsider effects needed to be considered. All members of the project team have considerable experience in the field of standard setting, ranging from senior research roles in exam boards (Baird and Gray), senior positions within regulators of qualifications and examinations (Gray, Isaacs and Opposs) and experience in university faculties (Baird and Isaacs); hence, they are both insiders and outsiders in the standard setting process (Mercer, 2007). In addition, all the project leads have conducted extensive international research in curriculum and assessment. This confers the benefits of credible experience and backgrounds that Shenton mentions (2004: 68), but also a need for care: 'the researchers need to be reflexive and disclose what they bring to a narrative' (Creswell and Miller, 2000: 126). To ensure that we achieved reflexivity, the project team met monthly and for an immersive writing week, sharing ideas and drafts of materials, and discussing and challenging these in a process of co-creation.

Some of our project participants were conducting their research from a locus fully inside the relevant examination body and had to spend considerable time convincing their key stakeholders to allow their participation in the project. A substantial amount of early work involved providing assurances about how the project would work and the kind of protection that would be in place for participants. A crucial issue for some jurisdictions was whether their systems would be judged to be less than best

practice; consent discussions with those participants focused on reassurance that the project did not aim to compare systems with each other in order to judge or rank them.

In some jurisdictions, national assessments are developed by academics in higher education, and this provides a degree of independence from government or commercial interests; project participants who worked for a higher education institution reported feeling fewer constraints sharing critical analysis of their own standard setting systems. For those jurisdictions where the project participant was a senior employee of an exam board or other governmental or commercial body, organizational contexts and employment conditions appeared to provide strong constraints, and the risk of potential biases was increased. As we have noted, for all of our project participants, whether academics or exam board personnel, even setting out the official position may have been difficult, and there may have been variations in how full a description could be provided in each country-specific account. Indeed, participants expressed concerns about transparency; for example, describing it as 'opening Pandora's Box'.

Arguably, those working in exam boards, in particular, are predisposed to risk-averse attitudes due to the nature of their organization's work. This can make it difficult for exam board researchers to share the results of their research, particularly with colleagues and stakeholders outside the organization. This was certainly true for a number of participants, and even some of the academics who felt that they were free from institutional or government pressure reported that they needed to be circumspect in what they said publicly. This issue proved to be a stumbling block for other potential participants, and one or two who expressed interest in the project were unable to take part. Indeed, most of the potential participants had to give these issues a lot of consideration in the early stages of deciding whether or not to be involved in the project. Some needed time to reflect on how open they could be, especially in discussing controversies or in placing themselves into debates in which alternative views of their policies may be put forward (or even seen to be legitimized): during consent interviews, words like 'delicate' were used. When recruiting participants, the project team took great care to discuss the nature of the critical reflection required for this research project. Ensuring that participants understood that this was required, and that alternative views of their system would also be sought and presented, was the main issue to be addressed in the initial consent process.

In the pilot project, we noted our status as insider–outsiders and the resultant constraints on our research. These constraints became even more pertinent for our project participants, most of whom were still serving as

senior officials in the examination system that they were describing and analysing. The participants entered the project with a commitment to study, question and test their own practices (as recommended in Stenhouse, 1975: 144): this desire was expressed in almost all of our initial consent interviews. We have already noted that the few project participants who worked within higher education appeared to feel less constrained in discussing the limitations of their own assessment and standard setting systems. However, exam board participants did report feeling constraints.

Exam boards are public institutions, known in their jurisdiction or nation, identified in the media and called to account through democratic structures. The only way to protect the confidentiality and ensure the anonymity of our exam board researchers and their organizations would have been to exclude so much descriptive detail about the system that the technical and theoretical analyses would have been rendered meaningless. However, it was part of the design of the project that different jurisdictions should share information with each other, and we brought the project participants together via a major international symposium in order for them to do this. Achieving the aims of the project required that, within the project, confidentiality should be breached. Internal confidentiality was not possible, then, but it was important that both the project team and all project participants provide each other with mutual reassurances of external confidentiality on sensitive issues. (For a discussion of ethical issues in internal and external engagement, see Floyd and Arthur, 2012.)

By opening and framing this research with a discussion of the insider research issues that have been captured in Chapter 3 of this volume, we characterized the symposium as a safe space in which participants could be completely frank and open about their own practices and ideas, and could expect others to respond respectfully, and with attention to personal and political sensitivities. Those participants who decided to take part in the project agreed to share, frankly and openly, information that in other contexts they may have been constrained to share with organizational or system outsiders. Of course, this does not mean that robust conceptual exchanges did not occur. These exchanges enacted the stances outlined in Chapters 1 and 14, with discussion around whether psychometrics should be viewed as the most technically sound assessment paradigm, and whether curriculum-based assessment lacked theoretical underpinning. Despite some strongly held views, it was clear that the symposium provided a forum for the project team and the case study authors to co-create concepts and develop understanding of these; reactions to the symposium, after the event, suggested that we had succeeded in creating with our colleagues

> a space where people can share views, be respected and take seriously the commitment to finding lines of consensus about what should be done to address questions of validity and legitimacy that might arise in regard to what they currently do (Kemmis *et al.*, 2014: 36).

The case study and thematic chapters of this volume delineate the views that were shared, the questions of validity and legitimacy that were addressed and the consensus that was reached. Figure 2.2 provides an outline of our data-collection methods.

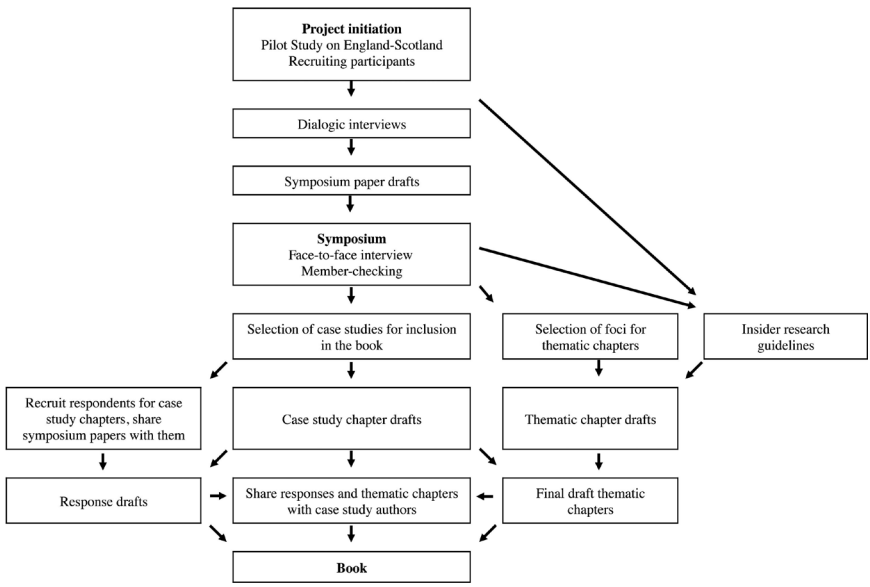


Figure 2.2: Research diagram

Using insiders as research leads and research participants necessarily brought both limitations and strengths; we aimed to offset the limitations and consolidate the strengths through careful research design, continual researcher reflection by the project team and the project participants and appropriate support and challenge for each other. However, we do not make truth claims for this data; instead we present our research as positional, 'acknowledging the inseparableness of the researcher and the process of inquiry' (Creswell and Miller, 2000: 129). This methodology chapter has outlined the steps that we took to provide assurances of authenticity and trustworthiness, but, ultimately, readers must judge for themselves.

References

Baird, J. and Coxell, A. (2009) 'Policy, latent error and systemic examination failures'. *CADMO*, 17 (2), 105–22.

Baird, J. and Gray, L. (2016) 'The meaning of curriculum-related examination standards in Scotland and England: A home–international comparison'. *Oxford Review of Education*, 42 (3), 266–84.

British Educational Research Association (2011) *Ethical Guidelines for Educational Research*. Online. https://goo.gl/AyKcU8 (accessed 7 June 2018).

Brinkmann, S. (2018) 'The interview'. In Denzin, N.K. and Lincoln, Y.S. (eds) *The SAGE Handbook of Qualitative Research*. 5th ed. Thousand Oaks, CA: SAGE Publications, 576–99.

Cohen, L., Manion, L. and Morrison, K. (2017) *Research Methods in Education*. 8th ed. London: Routledge.

Creswell, J. (1997) *Qualitative Inquiry and Research Design: Choosing among five traditions*. Thousand Oaks, CA: SAGE Publications.

Creswell, J.W. and Miller, D.L. (2000) 'Determining validity in qualitative inquiry'. *Theory into Practice*, 39 (3), 124–30.

Denscombe, M. (2010) *The Good Research Guide: For small-scale social research projects*. 4th ed. Maidenhead: Open University Press.

Floyd, A. and Arthur, L. (2012) 'Researching from within: External and internal ethical engagement'. *International Journal of Research and Method in Education*, 35 (2), 171–80.

Kemmis, S., McTaggart, R. and Nixon, R. (2014) *The Action Research Planner: Doing critical participatory action research*. Singapore: Springer.

Morse, J. (2018) 'Reframing rigor in qualitative inquiry'. In Denzin, N.K. and Lincoln, Y.S. (eds) *The SAGE Handbook of Qualitative Research*. 5th ed. Thousand Oaks, CA: SAGE Publications, 796–817.

Mercer, J. (2007) 'The challenges of insider research in educational institutions: Wielding a double-edged sword and resolving delicate dilemmas'. *Oxford Review of Education*, 33 (1), 1–17.

OECD (Organisation for Economic Co-operation and Development) (2015) *Education Policy Outlook 2015: Making reforms happen*. Paris: OECD Publishing.

Rossman, G.B. and Rallis, S.F. (2003) *Learning in the Field: An introduction to qualitative research*. 2nd ed. Thousand Oaks, CA: SAGE Publications.

Shenton, A.K. (2004) 'Strategies for ensuring trustworthiness in qualitative research projects'. *Education for Information*, 22, 63–75.

Sikes, P. and Potts, A. (eds) (2008) *Researching Education from the Inside: Investigations from within*. London: Routledge.

Silverman, D. (2011) *Interpreting Qualitative Data: A guide to the principles of qualitative research*. 4th ed. London: SAGE Publications.

Stenhouse, L. (1975) *An Introduction to Curriculum Research and Development*. London: Heinemann Educational.

Tracy, S.J. (2010) 'Qualitative quality: Eight "big-tent" criteria for excellent qualitative research'. *Qualitative Inquiry*, 16 (10), 837–51.

Wragg, E.C. (1994) 'Conducting and analysing interviews'. In Bennett, N., Glatter, R. and Levačić, R. (eds) *Improving Educational Management through Research and Consultancy*. London: Paul Chapman Publishing/Open University Press, 267–82.

Yin, R.K. (2014) *Case Study Research: Design and methods*. 5th ed. Thousand Oaks, CA: SAGE Publications.

Chapter 3

Researching national examination standards as an insider

Lena Gray

Introduction

As detailed in Chapter 2, we decided that expert insiders would be the best source of knowledge to investigate the methods and meanings of examination standards. We knew that choosing to approach senior insiders to be our project participants would bring methodological challenges: in the pilot project comparing England and Scotland we noted our own positions as insider–outsiders and the effects of this on our research. It was crucial to the project's aims that the participants' reports could be scrutinized for authenticity, but we needed to ensure that the project did not harm those who participated.

Insider research cannot involve objective observation and analysis; it is instead 'an encounter between individual choices and cultural tools employed in a particular institutional context' (Zembylas, 2003: 220). While some theorists would argue that this is the case for all social science research, insider research cannot but be situated in the researcher's own personal, organizational and political experience and context. Faced with charges of lack of objectivity, the insider researcher can defend themselves by building walls of data analysis, experimental technique and scientific method, or they can acknowledge that their own position is necessarily inextricable from the research that they are undertaking. Far from being a problem, this lack of objectivity means that insider research becomes an opportunity for the researcher to achieve authenticity in their research by being reflective and reflexive:

> Reflexivity suggests that researchers should acknowledge and disclose their own selves in the research, seeking to understand their part in, or influence on, the research. Rather than trying to eliminate researcher effects (which is impossible, as researchers

> are part of the world that they are investigating), researchers should hold themselves up to the light (Cohen *et al*., 2017: 303).

Senior professionals writing about systems inside their own organization face a number of problems in holding themselves up to the light. These problems are especially acute for professionals working inside exam boards. Chapter 2 described the highly political environments in which exam boards and their researchers find themselves; in such environments, organizations and individuals can be scapegoated, especially when a policy debate hits the media (McCaig, 2003). This can make it difficult for exam board researchers to share the results of their research, particularly with colleagues and stakeholders outside the organization.

The barriers facing exam board insider researchers are many and complex, and interact in ways that are unique to the field. Exam board researchers face the barrier of coming to their research with assumptions that they must try to unknow in order to be able to examine those assumptions reflexively and reflectively: they must struggle to avoid merely confirming their own beliefs. The researchers must shine a light, sometimes a cold one, on their own practices and the practices of their organization. This brings risks: the subject of their research can never be anonymous, and they will experience constant difficulties communicating about their research outside their organization. Like all insider researchers, they cannot leave the research field when their research is complete; they must continue an ongoing relationship with their research subjects. This situation can damage working relationships and make it difficult for the individual to communicate about their research inside their organization. In addition, they may find themselves directly or indirectly subject to organizational or governmental political pressures, or may even impose those pressures on themselves. Some of these issues are treated in the methodological literature, but rarely, if ever, in the combination of circumstances that affect exam board insiders. Thus, although some strategies have been identified that help individuals who work in such settings to exchange knowledge, work was needed to explore the particular combination of issues that could affect exam board insider research. This chapter discusses some of the barriers faced by exam board insider researchers and suggests ways that they can overcome those barriers to hold themselves up to the light.

Insiders, practitioners and researchers

Within qualitative social science research, awareness and discussion of insider research issues are increasing. The term ‘insider researcher’ is

itself a complex one, and can refer to a range of contrasting scenarios, including: professional staff carrying out research as part of a further qualification for career development purposes, staff whose day-to-day work includes responsibility for research among a range of more operational responsibilities, and staff whose job role is explicitly defined as that of 'researcher' (Sikes and Potts, 2008: 3–4). However, the insider researcher would define or name their role, the very fact of being an insider brings challenges:

> For the insider researcher who is also a 'proper' member of the setting they are investigating, the problem associated with criticisms around failure to maintain a distance in order to be able to take a clear and an unbiased non-partisan approach are significant and complicated. This is because adopting a distanced approach may, in some cases, be inimical to doing one's job in the way in which one has been hired to do it. People are expected to be loyal and committed to their employer and employing organisation and, while loyalty and commitment do not preclude taking an objective stance in order to develop and improve, detachment can be problematic in institutional terms (Sikes and Potts, 2008: 7).

To try to find solutions to those problems, we turned to the literature on participant observers. Participant observers are traditionally envisaged as researchers who enter a community under study in order to study it; they are part of the community only for the purposes of the research project (see, for example, Cohen *et al*., 2007: 404–8; Hammersley *et al*., 1994: 63–5; Denscombe, 2010: 206–15). This brings its own set of issues, which are extensively documented in the research literature (e.g. Cohen *et al*., 2017: 326–7; Maxwell and Beattie, 2004; Robson, 2002: 314–25). However, insider researchers do not enter their organizations in order to study them (Maxwell and Beattie, 2004). As Sikes and Potts (2008) noted, insider researchers are 'proper' members of the community they are researching; hence, they are not participant observers in the way the term is most commonly used in social science research. It would be more accurate to describe them as observing participants. We therefore turned to another body of methodological literature to try to conceptualize what this might mean in practice.

To help us understand the role of an observing participant, it is useful to consider the distinction between professional researchers and researching professionals or 'scholarly professionals' (Gregory, 1995: 182).

The latter are not employed primarily as researchers, and when they carry out research, they do it not for the research itself, but in order to develop their professional practice (Bourner *et al*., 2001: 71); in other words, to 'reflect on and illuminate their practice and the practice in the institution where they work' (Wellington and Sikes, 2006: 733). For the researching or scholarly professional, their research is not an end in itself but a means to professional or organizational development.

We have noted that as insiders, the key tool at the disposal of researching professionals is to reflect on their own practice, and so it is helpful to conceive of the job of the insider researcher as that of a 'reflective practitioner'. The idea of a reflective practitioner has its roots in the work of Dewey (1933) and Stenhouse (1975), and was probably most fully developed by Donald Schön (1983, reprinted 1991, 1987).

In his 1933 work, *How We Think*, Dewey set out to describe the kind of thought that has educational value, characterizing this as conscious, reasoned, sceptical and logical:

> *Active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusions to which it tends*, constitutes reflective thought (Dewey, 1933: 6, emphasis in original).

While arguing that reflective thought has value for educational purposes, Dewey provided examples that show us how such thought might operate in fields of professional endeavour or discovery. Describing how Columbus came to conclude that the world was round, Dewey declared:

> The thought of Columbus was a *reasoned conclusion*. It marked the close of study into facts, of scrutiny and revision of evidence, of working out the implications of various hypotheses, and of comparing these theoretical results with one another and with known facts. Because Columbus did not accept unhesitatingly the current traditional theory, because he doubted and enquired, he arrived at his thought (Dewey, 1933: 5–6).

Insider researchers need to be prepared for this kind of observation of and reflection on their own practices, and those of their organization; it involves close scrutiny of evidence, questioning of arguments and conclusions, and comparison of theories and processes as they reveal themselves through these practices. But while Dewey illuminated what insider researchers need to do, he did not tell us how to do it. For this, we turn to Stenhouse's

(1975) concept of extended professionalism. In the context of discussions on curriculum development, Stenhouse argued that curriculum could only be fully developed by teachers as researchers:

> The outstanding characteristics of the extended professional is a capacity for autonomous professional self-development through systematic self-study, through the study of the work of other teachers and through the testing of ideas by classroom research procedures (Stenhouse, 1975: 144).

Like Dewey, Stenhouse emphasized the importance of testing ideas through systematic reflection, but an additional element in the description of the extended professional is the notion of autonomous self-reflection: striving to gather evidence to allow evaluation of one's own practice.

Schön (1983) expanded on Stenhouse's arguments about the extended professional and set out a concept of 'reflection-in-action' as a means for professionals to develop their own knowledge and not be bounded by what he saw as the positivistic views of academic researchers. Schön's reflection-in-action is a way for professionals to deal with the complex, unique and inconsistent situations that they face. It is a way to increase knowledge in the face of undefined problematical situations: what he memorably calls the 'swampy lowland' of practice, in which messy, confusing problems defy technical solution' (Schön, 1983: 3). For Schön, professional practice is a 'complex, unstable, uncertain and conflictual' world (ibid.: 12). Others have seen this as a powerful way to deepen understanding of 'non-rational, unpredictable organizations' (Costley *et al*., 2010: 117):

> A significant advantage of the notion of the reflective practitioner is that it provides a conceptual framework within which the complexities, tensions and contradictions of work can be explored, and at the same time a reference point against which the intrinsic value of practice can be judged. The potential for practitioners to inform and influence policy, and the process by which they make considered responses to political, cultural and technological change and devise considered strategies to contain or exploit both intended and unintended consequences, are also key issues which are given prominence within a reflective practice model (Costley *et al*., 2010: 117).

Doncaster and Lester (2002) concluded, in their study about professional doctorate candidates using their research to develop 'capability', that reflective practice provides the practitioner with the tools to detach

themselves from the 'swampy lowland' (Doncaster and Lester, 2002: 100). For the exam board insider researcher, working in a field that is both technocratic and highly politicized, detaching themselves from that swampy lowland may seem like an impossible task. However, as Dewey reminded us in 1933, all reflective thinking is difficult:

> Reflective thinking is always more or less troublesome because it involves overcoming the inertia that inclines one to accept suggestions at their face value; it involves willingness to ensure a condition of mental unrest and disturbance. Reflective thinking, in short, means judgment suspended during further inquiry; and suspense is likely to be somewhat painful (Dewey, 1933: 13).

There is little doubt that insider researchers would agree that reflective thinking is troublesome and painful. Dewey, Stenhouse and Schön ask researching professionals to undergo that pain. Before we do, let us reflect on the nature of some of the unrest and disturbances that exam board insider researchers are likely to face.

Researching elites, elite researchers and confidentiality

Senior staff in exam boards are powerful people in the sense that they are 'those with great responsibility … whose decisions have significant effects on large numbers of people' (Cohen *et al*., 2007: 127). Their decisions affect many people and are subject to public scrutiny, and therefore any research involving them is sensitive because it is likely to encroach upon 'issues about which there is high-profile debate and contestation' (ibid.: 127).

While meanings of the term 'elite' may be open to dispute, senior exam board personnel fit the definition provided by Harvey (2011): 'those who occupy senior management and Board level positions within organisations' (433). Harvey also points to the 'significant decision-making influence within and outside of the firm' (ibid.: 433); in this context, senior exam board personnel could be said to be doubly elite – the organizations they work for undoubtedly have a high degree of influence on society and, in turn, are influenced and of interest to the public, policymakers and the media. Thus, as with the political elite investigated in several studies (see, for example, the range of studies reported in Walford, 1994), the natural inclination of senior exam board personnel may be to want to control and to resist transparency. For exam board insider researchers, this is likely to be in direct conflict with their inclination as researchers.

There are some interesting methodological texts that deal with researching the elite, but most are concerned with the interpersonal and

practical issues that may have to be overcome before, during and after an elite interview (see, for example, Berry, 2002; Conti and O'Neil, 2007; Selwyn, 2013); few texts have anything to say explicitly about the issues that may arise when powerful people are researching their own organization (Semel, 1994, provides one exception). Those texts that deal with researching bureaucratic elites seem particularly pertinent to the situation of exam board insider researchers. Harvey (2011), for example, documented the issues he faced interviewing elite subjects, noting that those who occupy senior decision-making positions are often scrutinized by journalists and therefore tend to want to control research about their work, seeing it as 'some form of challenge or justification for what they do' (Harvey, 2011: 433). Selwyn (2013) outlined his experience of how this works in the British senior civil service, discussing the rules and codes that individuals are subject to and the fact that it is written into their working arrangements that their role is to stay 'on message' (Selwyn, 2013: 342), meaning that they view any interview or research activity as an opportunity to create a 'rhetoric of justification' (ibid.: 342). Marshall (1984) noted that elite interviewees may obfuscate and avoid openness, even when information is already in the public domain. She provided a striking image for such behaviour:

> Some behave like ostriches. Scarred from past battles, investigations, and evaluations, they hide from any intrusion that might interrupt their orderly and secure existence (Marshall, 1984: 238).

If the bureaucratic elite prefers to stick its head in the sand rather than be open, what does this mean for exam board insider researchers, who are arguably part of that elite? Issues around internal and external confidentiality – and the difficulties of achieving these – are pertinent here. Much of the methodological advice on how to plan and conduct insider research, like research codes of practice and ethical guidelines more generally, stresses issues around the need to protect the confidentiality of participants (for example, Bell, 2005: 48–9; Blaxter *et al.*, 2006: 158–61; British Psychological Society, 2014: 9). For exam board researchers writing about practices in their own organizations, ideas of confidentiality and anonymity are irrelevant: they simply cannot be achieved. As Floyd and Arthur (2012) point out with regard to insider researchers in higher education, 'Whatever efforts are made to preserve anonymity, a simple online search will allow the most novice investigator to identify the institution' (Floyd and Arthur, 2013: 177).

Exam board insider researchers will often work for government bureaucracies or for organizations controlled or strongly influenced by such bureaucracies. In the Standard Setting Project, for example, most of the assessment organizations represented were either government departments, some kind of arm's-length agency of government, or under contract to government. In such a context, insider researchers may struggle to achieve openness and a sense of authenticity in their work. Even when our project participants were employed by universities and enjoyed academic freedom, the nature of the contractual arrangements under which they produced national assessments on behalf of government meant that some constraints were felt. If the exam board insider researcher is a senior member of that bureaucracy, they may themselves feel the need for obfuscation and justification, and may believe that anything else would exhibit disloyalty to their organization. Does this mean, then, that they find themselves, like Marshall's ostriches, with their heads in the sand?

The problem with adopting the ostrich position is that public trust in examinations is dependent, upon other things, on the public having a sense that the exam board is open and honest, although the technical nature of the work means that transparency is not necessarily the route to achieving this (O'Neill, 2002: 13–14, 2005: 18; Billington, 2007: 2; Newton, 2005: 76). The exam board insider researcher can be left with no route to transparency for their research and a feeling of being pulled in opposing directions. They want to be open for the sake of their research, but this openness may result in a decrease of public trust in their organization, and in them being perceived as disloyal to the organization they work for. How can the insider researcher resolve this dilemma? Is the only option to stick your head in the sand and hope you are invisible?

Lessons from action research: Communicative action

We turned to the methodological literature on action research in an attempt to find a resolution to the dilemma facing the exam board insider researcher. While not all insider research is action research, theories of action research were initially developed with particular reference to research within organizations (see Adelman, 1993), and later developments of those theories have stressed the socially situated nature of action research (see, for example, Kemmis, 1980). Both of these aspects suggest that action research theories may have much to offer the exam board insider researcher.

Action research is a form of insider research that has its roots in the work of Kurt Lewin in the 1940s (Kemmis, 1980; Adelman, 1993; Coghlan and Brannick, 2010). At its heart, action research is essentially research that

aims to bring about organizational change; it is as much about the change process as it is about the research. In its purest form, action research is also collaborative in nature: the researcher and the participants work together. It has gained much momentum in educational research because it has come to be seen as a way to help teachers to systematically reflect on and improve their own practice (Kemmis, 1980).

We found that theorists of action research had some useful lessons for the work of exam board insider researchers, especially those theorists who emphasize the collaborative nature of research. The most useful ideas draw on Jürgen Habermas's (1984) theories of communicative action and communicative spaces:

> I shall speak of *communicative* action whenever the actions of the agents involved are coordinated not through egocentric calculations of success but through acts of reaching understanding. In communicative action participants are not primarily oriented to their own individual successes; they pursue their individual goals under the condition that they can harmonize their plans of action on the basis of common situation definitions. In this respect the negotiation of definitions of the situation is an essential element of the interpretive accomplishments required for communicative action (Habermas, 1984: 285–6).

Habermas's theory of communicative action is a critical social theory that seeks to explain the social scientific project as essentially a linguistic activity, and one that involves agreement, negotiation, mutual understanding and consensus:

> The concept of communicative action refers to the interaction of at least two subjects capable of speech and action who establish interpersonal relations (whether by verbal or by extraverbal means). The actors seek to reach an understanding about the action situation and their plans of action in order to coordinate their actions by way of agreement. The central concept of interpretation refers in the first instance to negotiating definitions of the situation which admit of consensus (Habermas, 1984: 86).

Stephen Kemmis, one of the principal theorists of action research, sought a 'critical social science' that transcends subjectivism and objectivism (Kemmis, 1980), stressing the social aspect of research: such research would involve social questioning of and within a community, challenging collective understanding; the debate itself is the point of the activity. Drawing on

Habermas's work to develop these ideas, Kemmis *et al*. (2014) turned to the concept of the communicative space, describing it as involving 'a suspension of the strategic action we're ordinarily caught up in (getting things done), and an openness to rethinking what we are and could be doing' (Kemmis *et al*., 2014: 48). Suspending our strategic action and opening up communicative spaces provides room for the reflection and reflexivity – 'the constant analysis of one's own theoretical and methodological presuppositions' (Coghlan and Brannick, 2010: 41–2) – that are, as we have stressed, at the heart of insider research. Opening up such communicative spaces puts us in a strong position to carry out and communicate our insider research successfully.

Insider research guidelines

As we worked on the Standard Setting Project, we found ways to mitigate the risks of insider research by powerful people in public organizations and to open up communicative spaces that provide a place for reflection and reflexivity. In doing so, we realized the value of codifying the political and organizational barriers to such work and delineating a range of ways in which individuals and organizations could overcome them to advance their national examination technologies and policies.

We decided to produce guidelines, developed with the input of our project participants and other exam board insider researchers, to enable exam board researchers to be more transparent about the procedures they use and the challenges they face (Gray, 2017). The guidelines focus on how insider researchers can feel a sense of authenticity in their work. The document draws on the idea of 'speaking truth to power' (American Friends Service Committee, 1955), or the Foucauldian concept of 'parrhesia' (Foucault, 1983). The idea of parrhesia, which has its roots in ancient Greek philosophy and literature, implies speaking truthfully for the sake of common good – even when that is not recognized by the majority – and at considerable personal risk.

The guidelines support such activity and help exam board researchers to situate their research on a firmer methodological, conceptual and ethical basis by suggesting ways in which they could create for themselves a safe, communicative space in which to critically analyse their personal practice, their organizational practice and the dominant policy and cultural environment within their own national setting. The key issues addressed include: how exam board practitioners can safely make use of confidential data; how to ensure that insider research projects achieve maximum impact

with minimum harm; and how insider researchers can achieve authenticity in their research work, given the constraints that they face.

The guidelines can be found in Appendix B.

References

Adelman, C. (1993) 'Kurt Lewin and the origins of action research'. *Educational Action Research*, 1 (1), 7–24.

American Friends Service Committee (1955) *Speak Truth to Power: A Quaker search for an alternative to violence*. Online. http://quaker.org/legacy/sttp.html (accessed 19 June 2018).

Bell, J. (2005) *Doing Your Research Project: A guide for first-time researchers in education, health and social science*. 4th ed. Maidenhead: Open University Press.

Berry, J.M. (2002) 'Validity and reliability issues in elite interviewing'. *PS: Political Science and Politics*, 35 (4), 679–82.

Billington, L. (2007) *Public Trust and High Stakes Assessment*. Manchester: AQA Centre for Education Research and Practice. Online. https://goo.gl/6ePfua (accessed 8 June 2018).

Blaxter, L., Hughes, C. and Tight, M. (2006) *How to Research*. 3rd ed. Maidenhead: Open University Press.

Bourner, T., Bowden, R. and Laing, S. (2001) 'Professional doctorates in England'. *Studies in Higher Education*, 26 (1), 65–83.

British Psychological Society (2014) *Code of Human Research Ethics*. 2nd ed. Leicester: British Psychological Society. Online. https://goo.gl/93BCDA (accessed 19 June 2018).

Coghlan, D. and Brannick, T. (2010) *Doing Action Research in Your Own Organization*. 3rd ed. London: SAGE Publications.

Cohen, L., Manion, L. and Morrison, K. (2017) *Research Methods in Education*. 8th ed. London: Routledge.

Conti, J.A. and O'Neil, M. (2007) 'Studying power: Qualitative methods and the global elite'. *Qualitative Research*, 7 (1), 63–82.

Costley, C., Elliott, G. and Gibbs, P. (2010) *Doing Work Based Research: Approaches to enquiry for insider-researchers*. London: SAGE Publications.

Denscombe, M. (2010) *The Good Research Guide: For small-scale social research projects*. 4th ed. Maidenhead: Open University Press.

Dewey, J. (1933) *How We Think: A restatement of the relation of reflective thinking to the educative process*. Lexington, MA: D.C. Heath.

Doncaster, K. and Lester, S. (2002) 'Capability and its development: Experiences from a work-based doctorate'. *Studies in Higher Education*, 27 (1), 91–101.

Floyd, A. and Arthur, L. (2012) 'Researching from within: External and internal ethical engagement'. *International Journal of Research and Method in Education*, 35 (2), 171–80.

Foucault, M. (1983) 'Discourse and truth: The problematization of parrhesia. 6 lectures given by Michel Foucault at the University of California at Berkeley, Oct.–Nov. 1983'. Online. https://foucault.info/parrhesia (accessed 8 June 2018).

Gray, L. (2017) *Overcoming Political and Organisational Barriers to International Practitioner Collaboration on National Examination Research: Guidelines for insider researchers working in exam boards and other public organisations*. Oxford: Oxford University Centre for Educational Assessment. Online. https://goo.gl/MJVx8L (accessed 8 June 2018).

Gregory, M. (1995) 'Implications of the introduction of the Doctor of Education degree in British universities: Can the EdD reach parts the PhD cannot?'. *The Vocational Aspect of Education*, 47 (2), 177–88.

Habermas, J. (1984) *The Theory of Communicative Action*. London: Heinemann.

Hammersley, M., Gomm, R. and Woods, P. (1994) *MA in Educational Research Methods*. Milton Keynes: Open University Press.

Harvey, W.S. (2011) 'Strategies for conducting elite interviews'. *Qualitative Research*, 11 (4), 431–41.

Kemmis, S. (1980) 'Action research in retrospect and prospect'. Paper presented at the annual meeting of the Australian Association for Research in Education, Sydney, Australia, 6–9 November. Online. https://files.eric.ed.gov/fulltext/ED200560.pdf (accessed 8 June 2018).

Kemmis, S., McTaggart, R. and Nixon, R. (2014) *The Action Research Planner: Doing critical participatory action research*. Singapore: Springer.

Marshall, C. (1984) 'Elites, bureaucrats, ostriches, and pussycats: Managing research in policy settings'. *Anthropology and Education Quarterly*, 15 (3), 235–51.

Maxwell, G. and Beattie, D.R. (2004) 'The ethics of in-company research: An exploratory study'. *Journal of Business Ethics*, 52 (3), 243–56.

McCaig, C. (2003) 'School exams: Leavers in panic'. *Parliamentary Affairs*, 56 (3), 471–89.

Newton, P.E. (2005) 'The public understanding of measurement inaccuracy'. *British Educational Research Journal*, 31 (4), 419–42.

Newton, P.E. (2015) 'Ripping off the cloak of secrecy'. In Gray, L., Jackson, C. and Simmonds, L. (eds) *Examining Assessment: A compendium of abstracts taken from research conducted by AQA and predecessor bodies, published to mark the 40th anniversary of the AQA Research Committee*. Manchester: AQA Centre for Education Research and Practice, 70–8. Online. https://goo.gl/BUpo1P (accessed 8 June 2018).

O'Neill, O. (2002) *Reith Lectures 2002: A question of trust*. BBC Radio 4. Online. https://goo.gl/SXxUdu (accessed 8 June 2018).

O'Neill, O. (2005) *Assessment, Public Accountability and Trust*. Online. https://goo.gl/mQDYTz (accessed 8 June 2018).

Robson, C. (2002) *Real World Research: A resource for social scientists and practitioner-researchers*. 2nd ed. Oxford: Blackwell.

Schön, D.A. (1983) *The Reflective Practitioner: How professionals think in action*. London: Temple Smith.

Schön, D. (1987) *Educating the Reflective Practitioner*. San Francisco: Jossey-Bass.

Selwyn, N. (2013) 'Researching the once-powerful in education: The value of retrospective elite interviewing in education policy research'. *Journal of Education Policy*, 28 (3), 339–52.

Semel, S.F. (1994) 'Writing school history as a former participant: Problems in writing the history of an elite school'. In Walford, G. (ed.) *Researching the Powerful in Education*. London: UCL Press, 204–20.

Sikes, P. and Potts, A. (eds) (2008) *Researching Education from the Inside: Investigations from within*. London: Routledge.

Stenhouse, L. (1975) *An Introduction to Curriculum Research and Development*. London: Heinemann Educational.

Walford, G. (1994) *Researching the Powerful in Education*. London: UCL Press.

Wellington, J. and Sikes, P. (2006) '"A doctorate in a tight compartment": Why do students choose a professional doctorate and what impact does it have on their personal and professional lives?'. *Studies in Higher Education*, 31 (6), 723–34.

Zembylas, M. (2003) 'Emotions and teacher identity: A poststructural perspective'. *Teachers and Teaching: Theory and Practice*, 9 (3), 213–38.

Chapter 4

What is standard setting?

Dennis Opposs and Kristine Gorgen

Introduction

We explained in the previous chapters how we conducted our research. Before moving on to the case study chapters in the next part of this volume, we turn our attention to the key concept behind the project: standard setting. In this chapter we first clarify the term 'standard setting'. Different methods of setting standards are then classified and discussed. We relate the practice of combining different sources of evidence, or using both quantitative and qualitative data, that is common in educational assessment, to similar approaches used in the social sciences. Finally, the methods each jurisdiction uses to set standards in its national, school leaving or university entrance examinations are investigated.

What standards are being set?

To teachers, politicians and assessment experts, the word 'standards' has various and sometimes very diverse meanings. As Stobart points out in the context of England:

> This ambiguity [in the meaning of 'standards'] leads to the August ritual of any improvements in the GCSE/GCE pass rate being welcomed by some as an improvement in [performance] standards and denounced by others as further evidence of falling [examination] standards (as cited in McGaw *et al.*, 2004: 3).

It is therefore helpful at the outset to be clear about our meanings of different kinds of standards.

Content standards refer to the syllabus, curriculum or programme of study that sets out the content to be learnt or desired learning outcomes. They also prescribe what can be assessed. For example, a set of content standards may describe the specific knowledge, skills and understanding required of students studying for a particular examination in physics (Hambleton *et al.*, 2012). Content standards can be made more or less demanding by increasing or decreasing the breadth of material to be learnt, or the breadth of skills to be acquired. Content standards can also be made more or less demanding by increasing or decreasing the depth to which

the subject matter is studied, or the level of proficiency of the skills to be acquired.

In a similar sense, there are sometimes comments made about the standards of examinations that are really about the level of demand of the questions. Examinations of the same content can be made more or less demanding, for example, by adjusting the level of abstract thinking needed when students have to tackle a question. It is possible to set a highly demanding assessment, the content of which might be seen as low demand. Equally, there can be very demanding content with assessment that is low demand. Neither is likely to be good assessment.

In the sorts of examinations with which this book is mainly concerned – national, school leaving or university entrance examinations – there are typically several performance standards (or grade standards) set. Each of these performance standards is reported as a letter or number grade. In this context, the performance standard can be thought of as the minimum score required in an assessment for the student's responses to be sufficiently good to be labelled with that grade. The use of grades rather than scores is intended to assist users in making sense of the outcomes.

In the rather different setting of the workplace, there is usually a single pass-fail performance standard set in the assessments used for some occupations. This is the threshold standard which has to be reached in order to pass and thereby gain a licence to practise.

When we use the term 'standard setting' in this book, it is in the sense of setting performance standards. It is not about setting content standards. Neither is it about setting the level of demand of examinations.

Performance standards are often considered to be the most important aspect of an assessment system because of the uses to which they can be put. Linn (2003) describes four potentially important uses: exhortation; exemplification of goals; accountability for schools and teachers; and certification of student achievement. Many of the case studies presented in Chapters 5 to 13 of this volume discuss issues related to the use of assessment results in their jurisdictions.

Defining standard setting

> There is simply no way to escape making decisions. […] These decisions, by definition, create categories. If, for example, some students graduate from high school and others do not, a categorical decision has been made, even if a graduation test was not used. (The decisions were, presumably, made on *some* basis.)

> High school music teachers make decisions such as who should be first chair for the clarinets. College faculties make decisions to tenure (or not) their colleagues. We embrace decision making regarding who should be licensed to practice medicine. All of these kinds of decisions are unavoidable; each should be based on sound information; and the information should be combined in some deliberate, considered, defensible manner (Mehrens and Cizek, 2001: 478–9).

The examinations that have formed part of this project are typically those where the main purpose is to assist universities with making decisions about the right students for their courses. There are usually other purposes, too, such as supporting employers short-listing applicants for a job. To achieve that, the examinations provide, as an outcome for each student, a score or grade. Sometimes these scores or grades are aggregated across all the subjects examined; sometimes they are not.

The examinations themselves normally comprise various questions. The students' responses to these questions are marked, marks being allocated to each response according to its quality – to what extent it matches the expectations of the examiners about the correct answer. The marks are then aggregated. There may be further aggregation processes such as combining the outcomes from different papers, perhaps including school-based assessment results, possibly involving differential weights being applied to each.

The aggregated marks will usually be converted onto a separate scale that is used to report the results of the examination and will be intended to allow users of the results to interpret them more readily. This scale might use letter or number grades, or it may use scale scores. Where this process involves changing marks into grades, cut scores must be determined. Each grade then tallies with the marks between two adjacent cut scores.

In this book we use the term 'standard setting' to incorporate any process by which raw marks are converted into the reported outcome. This is a much broader definition than is common in the standard setting literature, which we briefly cover in the following section entitled 'Standard setting methods'.

In the psychometric literature, the term 'standard setting' is used to describe the process by which cut scores are set on the data from an examination to create categories used in reporting. Categories might be, for example, pass/fail, pass/merit/distinction, 1/2/3/4/5/U and A/B/C/D/U. In this context Cizek defined standard setting as 'the proper following of a

prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more states or degrees of performance' (Cizek, 1993: 100).

There is a separate concept in the psychometric literature that is called 'linking'. For two tests, a link between their scores is a transformation from a score on one to a score on the other (Holland and Dorans, 2006). In ideal circumstances, the linking can be described as equating. In test equating, a direct link is made between a score on one test and a score on another test so that the scores from each test can be used interchangeably. For test equating to be successful, several requirements have to be met: for example, the two tests should measure the same constructs and should have the same reliability. Equating allows a standard set judgementally on the first version of a test to be applied to subsequent versions using statistical methods rather than judgemental methods. In this book we include equating as a form of standard setting.

In several examination systems used as case studies in this project, the process followed to convert students' marks into reporting categories would better be described as maintaining standards. Typically, these systems have a relatively small number of reporting categories, normally letter grades. The aim in maintaining standards is to ensure that the standard of a grade in one examination is comparable with that issued when an earlier version of the same examination was taken, often one year earlier. In this book we use the term 'standard setting' to include both when the standard is being set for the first time and when it has previously been set and is now being maintained.

In other examination systems used as case studies in this project, the outcome reported uses a scale involving larger numbers rather than a reporting category. The scores reported are known as scale scores (Kolen, 2006). Normative information can be incorporated into a scale score to help users better understand their meaning. For example, by setting the mean scale score to be 200, users can understand whether a student is above or below the mean of those taking the test. In this book, this process too is taken to be a form of standard setting.

So from here on in, our definition of standard setting is any process by which raw marks are converted into the reported outcome.

Standard setting methods

Literature on the main standard setting methods

A large number of different standard setting methods are used in national, school leaving or university entrance examinations around the world.

Three of the most common judgemental methods – Angoff, bookmark and awarding – are described very briefly below. We also reference some other methods – see Table 4.2 in this chapter.

A widely used judgemental method to set cut scores is the Angoff method. In its commonest form, this requires members of a standard setting panel to review all the items that comprise an examination. (Often the panel members sit the examination to achieve this familiarization.) They then estimate for each item the probability that a borderline student – one on the cut score – taking the examination would provide the correct answer. The minimally acceptable score is then the aggregate of the probabilities. In practice, panel members spend some time making sure they are clear about the idea of a borderline student for each of the borderlines for which they have to make a judgement. Normally at least two rounds of ratings are carried out with opportunities for panel members to discuss their judgements and consider data between rounds (Thorndike *et al*., 1971; Hambleton and Pitoniak, 2006; Cizek and Bunch, 2007; Cizek, 2012).

Another method used widely is the bookmark procedure. Here the items that make up an examination are rearranged into a book with one item on each page. The pages are sequenced so that the items' empirical difficulty increases through the book. Panel members are then asked to identify the page where a borderline student will have a 0.67 probability of answering the item correctly. The average page number from the panel members' proposals is then used as the cut score (Hambleton and Pitoniak, 2006; Cizek and Bunch, 2007; Cizek, 2012).

A recent addition to the methods described in the US literature is the body of work method (Hambleton and Pitoniak, 2006; Cizek and Bunch, 2007; Kingston and Tiemann, 2012). This appears to be broadly similar to the arrangements originating in the UK which are called awarding. Awarding (or grading) is also a process by which the position of cut scores (known as grade boundary marks in this method) is determined. Awarding involves examinations that are not pre-tested and so is carried out after the students have sat the examination and their work has been marked. Panels of subject experts then consider both qualitative and quantitative information, including statistical data based on the actual marks obtained, and recommend a cut score for different grade boundaries (Robinson, 2007).

Kane's (2017) view that standard setting involves the development of policy statements about how good is good enough is applicable to all of the abovementioned methods. The results may be described as arbitrary in

the sense that there is no one right answer, but they can be reasonable, have acceptable consequences and be well supported by data.

A more statistical approach to standard setting than those methods described above involves making cohort-referenced assumptions. Typically, a target mean, standard deviation and range are set in advance of the examination being sat. Raw marks are then converted into a scale score using transformations (Kolen, 2006).

Since the 1980s, as the power of computers increased considerably, the use of a family of statistical models to analyse item data by means of item response theory (IRT) has become much more common. At the heart of each IRT model is a description of the probability that an examinee with particular characteristics will give a particular response to an individual item that has its own particular characteristics. Given that information, it is then assumed that responses for different items are conditionally independent. IRT can be used in scaling, equating, determining cut scores and score reporting (Yen and Fitzpatrick, 2006).

Over time, different attempts have been made to categorize standard setting methods. Cizek and Bunch (2007: 9–11) describe three categories, summarized in Table 4.1.

Table 4.1: Two-dimensional categorization schemes

Examinee-centred *v* *Test centred* (Jaeger, 1989)	*Examinee-centred*: judgements about whether real examinees show the necessary standard; could also be called 'holistic'.	*Test-centred*: each item or collection of items is considered and a judgement made of how a hypothetical examinee would perform.
Holistic models *v* *Analytic models* (Kane, 1994)	*Holistic models*: achievement or skill is assumed to be highly integrated.	*Analytic models*: achievement can be assessed using relatively small parts of performance.
Norm-referenced *v* *Criterion-referenced* (Christie and Forrest, 1981; Glaser, 1963/1994; Wiliam, 1996)	*Norm referenced*: performance standards are established with respect to the relative standing of examinees from a relevant population.	*Criterion-referenced*: performance standards are based on the level of knowledge, skill or ability necessary for a specified purpose and cut scores are established with respect to that level.

In practice, many of the methods used tend not to fall wholly into one category. So, for example, Angoff procedures might usually be classified as test-centred rather than examinee-centred, analytic rather than holistic and criterion- rather than cohort-referenced. Commonly, though, impact feedback – data that helps participants understand the consequences of their judgements on the population of examinees that are subject to the panel recommendations – is used in the process. Use of such data in the process muddies the categorization waters as it seems that independent criterion-referenced judgements cannot be made with confidence without first considering their implications for the overall results of the examination concerned. So decisions are made using both quantitative and qualitative evidence.

Cizek and Bunch (2007) concluded that while the two-dimensional categorizations presented in Table 4.1 are useful, 'the demands and nature of standard setting in practice compel us to conclude that no simple distinctions between methods can be made and that well-conceived and implemented standard setting must recognize that any procedure requires participants to rely on both dimensions to effectively carry out their task' (Cizek and Bunch, 2007: 10). Again, this indicates that typically, standard setting involves the use of both quantitative and qualitative sources of evidence.

The variety of methods that we consider in this volume to be part of standard setting techniques is wider than those considered by Cizek and Bunch (see Table 4.1 above) as we include statistical as well as judgemental techniques. One way of categorizing our wider set of methods is to consider whether or not they involve the use of experts to make judgements about the examinations themselves. If that judgement primarily involves making decisions about how well examinees might perform on individual items, then we call that an atomistic method. If the judgement makes major use of the quality of examinees' responses or their marks allocated when sitting a whole examination paper, then it is an aggregate method. Other methods that largely or wholly concern applying statistical techniques to the marks from students' responses we call statistical methods.

Some commonly used methods are categorized in Table 4.2 below. Readers interested in further details about these methods can find them in the references given. Some methods are also described in Chapters 5 to 13 of this volume.

Table 4.2: Standard setting methods

Atomistic methods Judgement about the examinations primarily involves making decisions about how well examinees might perform on individual items	Bookmark (Hambleton and Pitoniak, 2006; Cizek and Bunch, 2007; Cizek, 2012) Angoff and variations (Hambleton and Pitoniak, 2006; Cizek and Bunch, 2007; Cizek, 2012; see also Chapters 12 and 13) Direct Consensus (Hambleton and Pitoniak, 2006; Cizek, 2012) Nedelsky (Hambleton and Pitoniak, 2006; Cizek, 2012) Ebel (Hambleton and Pitoniak, 2006; Cizek, 2012)
Aggregate methods Judgement makes major use of the quality of examinees' responses or their marks allocated when sitting a whole examination paper	Contrasting Groups (Hambleton and Pitoniak, 2006; Cizek, 2012) Borderline Groups (Hambleton and Pitoniak, 2006; Cizek, 2012) Body of Work (Hambleton and Pitoniak, 2006; Cizek and Bunch, 2007; Cizek, 2012) Awarding (Robinson, 2007) See also Chapters 6 and 11
Statistical methods Methods which largely or wholly apply statistical techniques to the marks from students' responses	Item Response Theory (Yen and Fitzpatrick, 2006) Norming (Kolen, 2006) Scaling (Kolen, 2006) See also Chapters 5 and 8

The atomistic and aggregate methods combine, to varying degrees, quantitative and qualitative sources of evidence to decide on the quality and reported outcome of a student's work. Although methods for setting cut scores have been a consistent focus of attention over the years (see, for example, Newton, 2005), it has been recognized that exactly how evidence is combined to enable decisions to be made during standard setting has largely been ignored (Newton, 2000: 40). We now turn to literature on mixed methods design to help conceptualize how different kinds of data are amalgamated when setting performance standards using atomistic and aggregate methods.

Combining sources of evidence

The practice of combining both quantitative and qualitative sources of evidence to arrive at an answer or conclusion is not unique to the practice of standard setting. In social sciences, mixed methods research techniques involve the connection, integration or linking of two independent strands of quantitative and qualitative data. In this century, interest in mixed method design has risen. Morse (2010) stated that researchers have seen that in mixed method design, quantitative and qualitative designs, which have been at odds for decades, may be able to exist together. We suggest that the same holds true for standard setting methods, in which combining sources of evidence creates an outcome that is controlled, rigorous and complex (Morse, 2010).

As in any kind of scientific research, the choice to combine quantitative and qualitative methods in a mixed methods research design should be based on the specific aims and interests of the research project (Creswell, 2009; Punch, 2005; Ridenour and Newman, 2008). If we apply this to the context of setting standards in national examinations, it might be government policy that defines the aims of the standard setting process. The research design, so the method(s) used to set standards, would then be chosen to enable the standards to be set and maintained in line with that given aim.

How quantitative and qualitative sources of evidence can be combined in standard setting will be illustrated with the example of A level examinations in England. The government policy that underlies standard setting in England is to keep standards comparable over time. Ofqual implements this policy through its adoption of the comparable outcomes approach. Examination boards are then required to follow the policy when they are setting grade boundaries for their own examinations. The term 'comparable outcomes' deserves more attention here, since it permeates the standard setting process in England. As discussed in more detail by Taylor and Opposs in Chapter 6, Ofqual's comparable outcomes approach is based on the assumption that if the cohort taking the examinations this year is similar in size, background and experiences to last year's cohort, then results should be similar. According to Newton (2011), Ofqual's assumption that student outcomes should be comparable over the years has long been championed by English examination boards and is used as a 'rule-of-thumb' that shapes the sources of evidence and methods used in the awarding process (Newton, 2011: 23).

The sources of evidence and their use in the standard setting process in England are specified in Ofqual's now-obsolete *Code of Practice* (2011). Distinguishing between quantitative and qualitative evidence, the *Code of Practice* states that 'certain types of evidence will be more appropriate when maintaining qualification standards over time than when setting standards in a new qualification' (p. 40). While quantitative and qualitative evidence might be weighted differently depending on the subject, both types of evidence are considered, which makes the standard setting process in England a mixed methods design. As Taylor and Opposs point out in Chapter 6, the 2011 *Code of Practice* has been withdrawn and the examination boards are no longer required by Ofqual to abide by it. However, they still tend to follow the procedures described in the *Code* and use both quantitative and qualitative evidence when maintaining standards.

How quantitative and qualitative data are used in scientific research is explained by Tashakkori and Teddlie (2010). Mixed methods are therefore often employed when both validity and credibility are sought, as is the case with setting examination standards.

Five design elements are important when considering how data or sources of evidence are combined in mixed method designs (based on Tashakkori and Teddlie, 2010). Figure 4.1 below shows these design elements as applied to the example of A level examinations in England.

1) Which theoretical drive underlies the design?
   The theoretical drive defines the research design. It is concerned with the logical reasoning and necessary evidence to support that reasoning. If the main theoretical drive is inductive, the research design is usually more qualitative. If the main theoretical drive is deductive, the design is usually more quantitative.
2) What is the core component (major method for collecting data)?
   Depending on the theoretical drive, evidence/data would either be collected using quantitative or qualitative methods. Quantitative methods use statistics and require a large quantity of numerical data. Qualitative methods often use smaller sample sizes and are frequently employed to increase the depth of analysis.
3) What is the supplemental component (additional data collection method)?
   It is possible to have the same kind of method (qualitative or quantitative) in both the core and supplemental components. This can occur when different types of quantitative or qualitative data are used or when one source uses micro data and is combined with macro data of the same

kind. However, it is more common to supplement the core component with the alternative method.

4) What is the order in which the two components are used for data collection?
   Research designs can be simultaneous or sequential: simultaneous designs are those in which both components are considered side by side; sequential designs are usually used when one component builds on the previous component.
5) What is the point of interface at which the components' results are combined?
   If the supplemental component's results are included in the analysis stage of the core component, it is called an analytic point of interface. If the two components are consolidated at the stage of presenting the results it is called a results point of interface.

Figure 4.1 also shows how England's A level standard setting process can be interpreted as a mixed methods design with a quantitative core component and a qualitative supplemental component. The theoretical drive underlying the design is the comparable outcomes approach based on the premise that, all other things being equal, outcomes should be similar across different cohorts. The core component is quantitative, as statistical evidence is more heavily weighted in England's standard setting process (at least when there is a sufficiently large number of students to provide reliable statistical predictions). Qualitative evidence, such as students' responses to examination questions or to school-based assessment tasks and the awarders' judgements of those responses, is used to support the statistical data. The design is simultaneous as both quantitative and qualitative evidence is used and combined in each stage of the process. The standard setting process has an analytic point of interface, as the two kinds of evidence are simultaneously considered at each stage.

The next part of this chapter will turn to other jurisdictions involved in the Setting Standards Project, discussing which methods they each use when setting standards.

Examples of standard setting in different jurisdictions

The Standards Setting Project involved 12 jurisdictions, each of which uses and combines methods differently when setting standards. In this section we provide short summaries of those standard setting processes. The examination systems in nine of the jurisdictions also feature in Part Two of this volume as case studies.

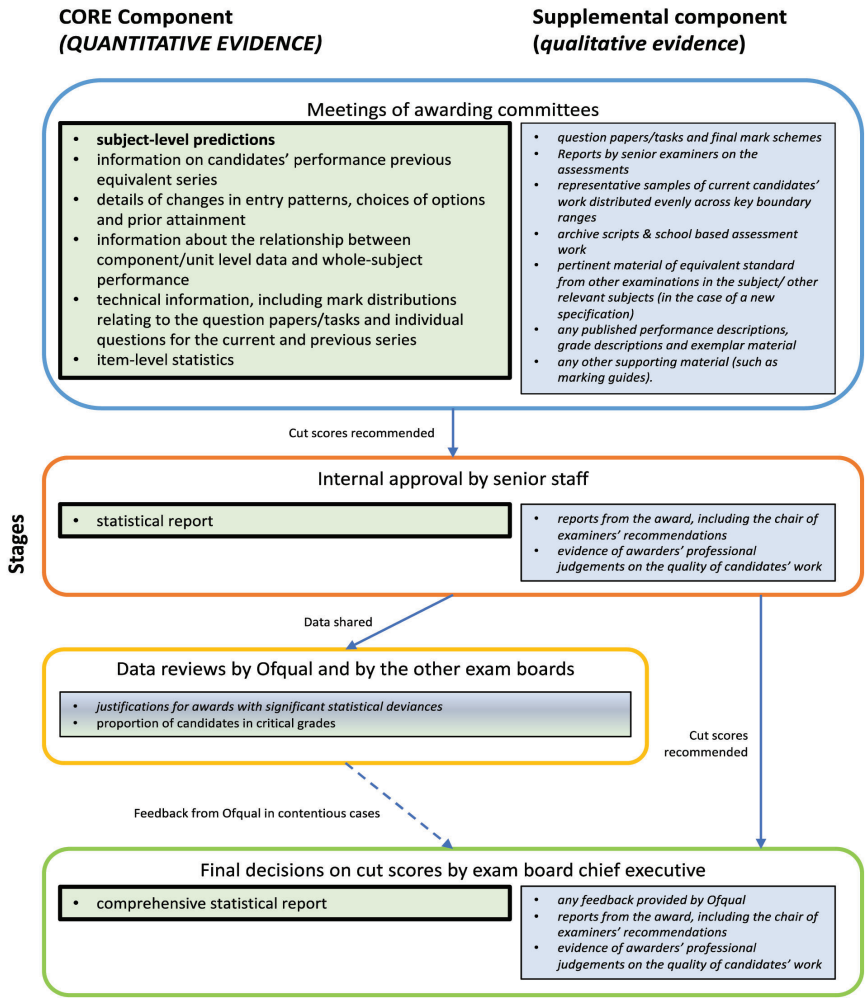


Figure 4.1: England's A level standard setting process

The summaries are based on data provided by the participants in the project from each of those jurisdictions. In the paragraphs that follow we have tried to classify how jurisdictions combine sources of evidence in their standard setting processes. The core component is written in upper-case letters (QUAN or QUAL) and the supplemental component in lower-case letters (quan or qual). The upper-case designation comes first unless the sequencing of the process has the supplemental component first.

In 8 of the 12 jurisdictions, one key intention of the standard setting process employed is to maintain over time the performance standards of the grades. A student given a grade A based on the 2018 examinations

should be showing broadly the same level of performance as a student awarded a grade A in 2017. The eight systems use letter or number grades to report results. England, France, Hong Kong, Ireland and South Africa each use aggregate methods to set standards – evidence from the quality of examinees' responses or examinees' marks from a whole examination paper playing an important part. Sweden and Queensland rely on teachers' judgements of their students' work to determine grades. The Advanced Placement examinations in the US use a modified Angoff technique to set standards – an atomistic method.

Hong Kong

The standards setting process used in Hong Kong's Diploma of Secondary Education Examination is described as being standards-referenced reporting. Standards-referenced reporting aims to report students' results against a set of prescribed levels of achievement based on typical performances of candidates at those levels. The results are expressed in terms of five pass levels from 5 (the highest) to 1. At the top, Grade 5 is divided. The best 10 per cent of Grade 5 performances are awarded a 5**, and the next 30 per cent a 5*. Descriptors explain what the typical candidate performing at each level is able to do. These descriptors are important reference sources for subject experts to make judgements when setting standards (known as grading).

The performance standards of levels 5 to 1 of the four core subjects (Chinese language, English language, mathematics and liberal studies) were set in the first year of the examinations (2012) using expert judgement. The standards have been maintained in the years since 2013 using various statistical data and reference to candidates' current and past levels of performance as well as expert judgement. In particular, a monitoring test for the four core subjects is carried out each year and a Rasch model is used to produce recommendations for cut scores for the consideration of the expert panels.

The standards setting process in use aims to ensure that no single factor or subject expert can predominate in the decision making, and the standards can be maintained and held constant without any 'grade inflation' over time. This is a mixed methods design as it uses both statistical predictions and expert judgements of students' work. The core component here is quantitative evidence; the supplemental component is qualitative evidence (QUANqual).

England

In England's A levels, the six pass grades that students can achieve are A* (the highest) through A to E (for further information about this system, see Chapter 6). The basic principle behind the standards maintaining process (known as awarding) is to retain from year to year the level of performance at a grade boundary mark. To help achieve this, the examination boards draw on both statistical and judgemental techniques. The key statistic used at the awarding meeting is a prediction based on prior attainment. The predictions map the relationship between prior attainment and A level outcomes for students taking each subject in a reference year. The examination boards use this relationship to predict the outcomes for the current cohort of students based on their prior attainment. If the prior attainment of the current cohort remains similar to that of the previous cohort, then the outcomes would be expected to be similar.

The awarders – senior examiners – scrutinize the students' examination work (called scripts) around the predicted grade boundary marks (cut scores), comparing them with the quality of scripts from the same grade boundaries from the previous year (called archive scripts), before using their judgement to recommend the grade boundary marks which are then applied to all students. As described in Figure 4.1, this is a mixed methods design using both statistical predictions and expert judgements of students' work (QUANqual).

So in England and Hong Kong, the cut scores are adjusted from year to year with the aim of ensuring that the standard of performance associated with each grade remains consistent over time. Ireland's Leaving Certificate and South Africa's National Senior Certificate achieve the same aim through a different approach.

Ireland

The State Examination Commission (SEC) in Ireland describes its Leaving Certificate examination as attainment-referenced (further information about the Irish system is given in Chapter 9). Since 2017 it has used a grade scale running from 1 (the highest grade) to 8. Each grade corresponds in a predetermined way to a percentage range of the marks obtained. So a grade 4, for example, always relates to a mark range of 60–69 per cent.

The mark therefore determines the grade in a pre-ordained fashion that is fixed over time and across subjects. This poses considerable challenges for maintaining consistency in grading standards over time, since it is impossible to guarantee that a particular year's examination questions will be identical in demand to those used in any other year.

To solve this problem, a standard setting process is embedded within the marking process. If there are indications that marking is producing a grade distribution considered inappropriate in the context of statistics from previous years and the levels of achievement being observed, adjustments to the mark schemes are used to achieve changes in the distribution of the raw marks and hence the grades.

The linking process in Ireland uses fewer sources of information than is the case in the awarding process in England. Scripts from the same grade boundary from previous years are not generally used for comparison. The senior examiners make judgements of students' work based on their knowledge and experience of examination standards. Changes to the size of the cohort are considered when evaluating the emerging grade distribution but prior attainment data are not available. The 'similar cohort adage' (Newton, 2011: 22) is a dominant influence; if the cohort is large, aligning grade boundary standards across different examinations can best be achieved by mainly using statistics. In Ireland, expert judgement is used as a check rather than as the main control. Again, this is a mixed methods design where the core component is quantitative – the use of statistics. The qualitative judgement of students' work is the supplemental component (QUANqual).

South Africa

South Africa's National Senior Certificate adopts a similar approach to Ireland in maintaining grading standards over time (for further information about this system see Chapter 11). Results are reported on a scale running from 7 (the highest) to 1. Each grade corresponds in a predetermined way to a percentage range of the marks obtained. So a grade 6, for example, always relates to a mark range of 70–79 per cent.

Mark distributions for the current examination and the corresponding average distributions over a number of years are compared to determine the extent to which they correspond. If there is good correspondence, in terms of the mark distribution statistics and pass rates, then it can be accepted that the examinations were of equivalent standard and no changes are made.

If there are significant differences, then attempts are made to ascertain the reasons for those differences. There may, for example, have been a clear change in the composition of the group of students taking a particular subject. In the absence of strong indications of valid reasons for differences, it is generally accepted that the differences are due to deviations in the demands of the examination or in the marking, and the marks are adjusted to compensate for these deviations. This is another mixed methods

design where the core component involves the use of statistics and the qualitative evidence in the form of subject reports, which are used only if the statistics show significant differences over time, is the supplemental component (QUANqual).

France

In the baccalauréat in France, each subject uses the same marking scale, with marks from 0/20 up to 20/20 (with the possibility of half and quarter marks). To pass the baccalauréat a student must have a mark aggregate average of at least 10 out of 20. This is another examination system that aims to maintain the performance standard over time, and the mechanism for achieving that is embedded in the marking system. However, no statistical methods, such as those described in the examples above, are used in this mechanism (further information about the French baccalauréat is given in Chapter 7).

A mark scheme (used in most but not all subjects) describes what mark should be allocated to different questions and the answers expected. For the subjects where no mark scheme is provided, the expectations are implicit and should be part of the professional expertise of the teachers.

The other support provided to ensure that marking is accurate is the existence of *commissions d'harmonisation*, one for each subject. These are groups of experienced teachers and inspectors from the local level, who will join markers during the marking process in order to help adjust and, to some extent, standardize the marking.

The baccalauréat does not appear to be a mixed methods design. There is no quantitative evidence used in the process at all.

Sweden

In Sweden, students' grades are determined by their teachers using different sources of evidence (for further information about the Swedish system, see Chapter 12). Typically, teachers use a type of portfolio-approach in which course work, teacher observations (notes) and national test scores are combined to give a composite grade. Individual teachers decide how to weight each element. The pass grades are on a scale running from A (the highest grade) to E.

The standard setting is a particularly important step in the development phase for the national tests since the cut scores are determined before the tests are administered. This is to prevent teachers interpreting test scores in a relative fashion.

Given the requirement for cut scores to be fixed before students take the tests and the inclusion in the tests of both dichotomously and polytomously

scored items, the modified Angoff method is used in establishing the cut scores. Whole cohort performance statistics that are so important to many systems are not available here. Instead, in the final stage of determining cut scores for the national tests, the Swedes use item data from field testing. The Swedish arrangements for Angoff standard setting follows the approach recommended in the literature except for one alteration: it does not include a separate step for the determination of performance level descriptors. For the parts in the Swedish and English tests where the students write essays, the common standard setting method is the bookmark method.

The Swedish system is also a mixed methods design but the core component here is qualitative evidence – teachers' judgements of their students' work – and the supplemental component is quantitative evidence (QUALquan).

US

The Advanced Placement (AP) examinations in the US also use a modified Angoff method to set performance standards. Results are reported on a scale running from 5 (the highest grade) to 1 (for further information about this system, see Chapter 13).

The AP standard setting process, in the meaning that phrase has in the psychometric literature, involves panel-based expert judgement. Once they have been trained in the process, the subject matter experts use their knowledge and experience to provide two rounds of ratings, but there is a wish to have some connection with student outcomes. As standards are set on the AP examinations after the students are assessed, data from that administration are used as impact data after the first round of judgements. This is another QUALquan mixed methods design for standard setting. After the performance standards have been set for an examination, they are maintained in subsequent years through equating without the use of qualitative judgements.

Queensland

In Queensland's current system of externally moderated school-based assessment, all assessment is standards-based. Teachers make judgements about the quality of student achievement with reference to performance descriptions that describe how well students have achieved the objectives in syllabi. Within the syllabus for each subject, objectives are grouped by dimensions and presented in a standards matrix, which describes the standards for each dimension, expressed on a grading scale running from A (highest grade) to E. So this is not a mixed methods design as it uses

only qualitative evidence (for further information about this system, see Chapter 10).

In the reforms being introduced for students entering Year 11 in 2019, the assessment system will include external assessment. These comprise assessment tasks that are externally set and marked, focused on particular units or aspects of study. They are not necessarily terminal examinations assessing the full course of study. Final subject results for general subjects will be derived from a combination of three school-based assessments and one external assessment. The results across the four assessment tasks will not be scaled against one another but will instead be combined to provide an overall result. In this way, the assessment decisions of teachers will take priority over the results from external assessments. Final results in general subjects will be reported to students as a numerical result out of 100, with achievement of standards presented on an A to E scale. Queensland has a mixed methods design with the qualitative aspect predominating so can be described as QUALquan.

In 4 of the 12 case studies – Chile, South Korea, Victoria and Georgia – a more statistical approach is used. The maintenance over time of performance standards is not a primary concern. These systems typically use scale scores to report results.

Chile

The results of each administration of Chile's University Selection Test (PSU) are cohort-referenced (for further information about the PSU system, see Chapter 5). The PSU comprises four examinations with 80 multiple choice questions in each. The different forms of the examinations are equated and the final score estimated using the number of correct responses per student. The mark distribution is then normalized so that it has a mean of 500 points and a standard deviation of 110. For the normalization, the minimum and maximum score a student can obtain are set to 150 and 850 points, respectively.

Scores are not strictly comparable over time. Each year, each university uses its own criteria and experience to set the minimum scores required in its selection process. This does not appear to be a mixed methods design as no qualitative evidence is used in the process.

South Korea

In South Korea's College Scholastic Ability Test (CSAT), all subjects except two are cohort-referenced. Three results are reported for each subject: a standard score, a percentile rank and a level.

The standard score is calculated using a linear transformation method. Language arts, mathematics and English have a mean of 100, a standard deviation of 20 and a range of 0–200. Other subjects use a mean of 50, a standard deviation of 10 and a range of 0–100. The percentile rank indicates the percentage of students who fall below the midpoint of the given score interval. Levels, ranging from 1 (the highest) to 9, are determined based on students' standard score. So the top 4 per cent of students are in level 1, the next 7 per cent in level 2, the next 12 per cent in level 3 and so on.

CSAT is cohort-referenced, and no equating process is used to link the results from different examinations although test developers do try to maintain the same mean and standard deviation for each test over time. As with Chile, this does not appear to be a mixed methods design as only quantitative evidence is used in the process.

Victoria

Victoria's Certificate of Education (VCE) produces two reported outcomes for each individual student. One is a Study Score and the other is a letter grade. With respect to standard setting processes, these two reported outcomes are treated quite differently, though both share normative underpinnings.

Each VCE study (or subject) consists of up to four units, with each unit nominally delivered over one semester. Units 1 and 2 are usually undertaken in the penultimate year of senior secondary schooling and need not be taken in sequence. In the final year of a given study, Units 3 and 4 are undertaken in sequence. Study Scores, which are the final subject results for each Unit 3 and 4 sequence, are calculated by ranking students on the basis of their graded assessment scores from these two units. These rankings are then converted into a normal distribution of scores with a mean of 30 and a standard deviation of 7, truncated to range from 0 to 50. The standards associated with certain subject results are not immediately comparable with results for other cohorts in other subjects or in the same subject in other calendar years. Nevertheless, the assessment system that underpins the VCE is based upon the core assumption that student achievement is normally distributed to a greater or lesser degree; hence, within each study the students' results are more or less similarly distributed. This is another process that uses only quantitative evidence. Study Scores have greater status being the final subject result for each student, whereas letter grades assume a more descriptive role in student reporting.

Georgia

In Georgia's Unified National Examinations (UNE) all students take three mandatory examinations: Georgian language, a foreign language and a general aptitude test (GAT). Some students also take an additional field-specific examination (further information about the system in Georgia is described in Chapter 8).

After scoring is complete, raw scores are converted into the scaled scores that are reported. As there are usually multiple versions of each examination, scores across different versions of the same examination are first equated using percentile rankings. Then the scores are standardized to make different subject examinations comparable using the mean scores of each subject examination. Passing scores in all examinations are set just above the score an applicant would obtain by guessing closed-ended question responses randomly. Again, this is a process that uses only quantitative evidence.

Kane's view given earlier in this chapter is that standard setting involves the development of policy statements about how good is good enough (Kane, 2017). The case study chapters that follow present various understandings and operationalizations of good enough. Chapter 15 of this volume will then discuss in more detail why and how the concept of good enough is highly context specific and constantly evolving.

Conclusion

In this chapter we have built on the different meanings of standards found in the academic literature and how they relate to the definitions underlying the examination systems in different jurisdictions as described in Chapter 14. We have explained some of the meanings of the word standards and defined standard setting in a broad way to encompass any process where raw marks are converted into reported outcomes such as grades or scaled scores. We have then briefly described some commonly used standard setting methods, providing references for others. Ways of categorizing those standard setting systems have also been considered.

We have explained that most standard setting methods combine quantitative and qualitative sources of evidence. For the first time we have related that to the social science literature on mixed methods design. That allowed us, when describing the standard setting methods used in 12 different systems around the world, to categorize each of those systems in terms of the type of evidence they use to make standard setting decisions (see Table 4.3).

Table 4.3: Standard setting designs in 12 jurisdictions

Quantitative core component with a qualitative supplemental component (QUANqual mixed methods design)	England Hong Kong Ireland South Africa
Only **quantitative** evidence	Chile Georgia South Korea Victoria
Qualitative core component with a quantitative supplemental component (QUALquan mixed methods design)	Queensland Sweden US AP
Only **qualitative** evidence	France

Part Two of this volume offers a more detailed discussion and insider perspective on standard setting systems around the world.

References

Christie, T. and Forrest, G.M. (1981) *Defining Public Examination Standards* (Schools Council Research Studies). Basingstoke: Macmillan Education.

Cizek, G.J. (1993) 'Reconsidering standards and criteria'. *Journal of Educational Measurement*, 30 (2), 93–106.

Cizek, G.J. and Bunch, M.B. (2007) *Standard Setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.

Cizek, G.J. (ed.) (2012) *Setting Performance Standards: Foundations, methods, and innovations*. 2nd ed. New York: Routledge.

Creswell, J.W. (2009) *Research Design: Qualitative, quantitative, and mixed methods approaches*. 3rd ed. Thousand Oaks, CA: SAGE Publications.

Glaser, R. (1994) 'Instructional technology and the measurement of learning outcomes: Some questions'. *Educational Measurement: Issues and Practice*, 13 (4), 6–8.

Hambleton, R.K. and Pitoniak, M.J. (2006) 'Setting performance standards'. In Brennan, R.L. (ed.) *Educational Measurement*. 4th ed. Westport, CT: Praeger, 433–70.

Hambleton, R.K., Pitoniak, M.J. and Copella, J.M. (2012) 'Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results'. In Cizek, G.J. (ed.) *Setting Performance Standards: Foundations, methods, and innovations*. New York: Routledge, 47–76.

Holland, P.W. and Dorans, N.J. (2006) 'Linking and equating'. In Brennan, R.L. (ed.) *Educational Measurement*. 4th ed. Westport, CT: Praeger, 187–220.

Jaeger, R.M. (1989) 'Certification of student competence'. In Linn, R.L. (ed.) *Educational Measurement*. 3rd ed. New York: American Council on Education/ Macmillan, 485–514.

Kane, M. (1994) 'Examinee-centered vs. task-centered standard setting'. Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Kane, M. (2017) 'Using empirical results to validate performance standards'. In Blömeke, S. and Gustafsson, J.-E. (eds) *Standard Setting in Education: The Nordic countries in an international perspective*. Cham: Springer, 11–29.

Kingston, N.M. and Tiemann, G.C. (2012) 'Setting performance standards on complex assessments: The body of work method'. In Cizek, G.J. (ed.) *Setting Performance Standards: Foundations, methods, and innovations*. 2nd ed. New York: Routledge, 201–24.

Kolen, M.J. (2006) 'Scaling and Norming'. In Brennan, R.L. (ed) *Educational Measurement*. 4th ed. Westport, CT: Praeger, 155–86.

Linn, R.L. (2003) 'Performance standards: Utility for different uses of assessments'. *Education Policy Analysis Archives*, 11, 31. Online. https://goo.gl/FdvgaJ (accessed 8 June 2018).

McGaw, B., Gipps, C. and Godber, R. (2004) *Examination Standards: Report of the independent committee to QCA*. London: QCA. Online. https://goo.gl/qSxMWM (accessed 8 June 2018).

Mehrens, W.A. and Cizek, G.J. (2001) 'Standard setting and the public good: Benefits accrued and anticipated'. In Cizek, G.J. (ed.) *Setting Performance Standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, 477–85.

Morse, J. (2010) 'Procedures and practice of mixed method design: Maintaining control, rigor, and complexity'. In Tashakkori, A. and Teddlie, C. (eds) *SAGE Handbook of Mixed Methods in Social and Behavioral Research*. 2nd ed. Thousand Oaks, CA: SAGE Publications, 339–52.

Newton, P.E. (2000) *Maintaining Standards over Time in National Curriculum English and Science Tests at Key Stage Two: A report for the Qualifications and Curriculum Authority*. Slough: NFER. Online. https://goo.gl/EAAvPf (accessed 8 June 2018).

Newton, P.E. (2005) 'Examination standards and the limits of linking'. *Assessment in Education: Principles, Policy & Practice*, 12 (2), 105–23.

Newton, P.E. (2011) 'A level pass rates and the enduring myth of norm-referencing'. *Research Matters*, Special Issue 2, 20–6.

Office of Qualifications and Examination Regulation (2011) *GCSE, GCE, Principal Learning and Project Code of Practice*. Coventry: Ofqual. Online. www.gov.uk/government/publications/gcse-gce-principal-learning-and-project-code-of-practice (accessed 7 August 2018).

Punch, K. (2005) *Introduction to Social Research: Quantitative and qualitative approaches*. 2nd ed. London: SAGE Publications.

Ridenour, C. and Newman, I. (2008) *Mixed Methods Research: Exploring the interactive continuum*. Carbondale, IL: SIU Press.

Robinson, C. (2007) 'Awarding examination grades: Current processes and their evolution'. In Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority, 97–123. Online. https://goo.gl/rSVQgr (accessed 8 June 2018).

Tashakkori, A. and Teddlie, C. (eds) (2010) *SAGE Handbook of Mixed Methods in Social and Behavioral Research*. 2nd ed. Thousand Oaks, CA: SAGE Publications.

Thorndike, R., Angoff, W. and Lindquist, E. (1971) *Educational Measurement*. 2nd ed. Washington, DC: American Council on Education.

Wiliam, D. (1996) 'Standards in examinations: A matter of trust?'. *Curriculum Journal*, 7 (3), 293–306.

Yen, W.M. and Fitzpatrick, A.R. (2006) 'Item response theory'. In Brennan, R.L. (ed.) *Educational Measurement*. 4th ed. Westport, CT: Praeger, 111–53.

Part Two

Case studies

2

Chapter 5

Standard setting in Chile: The Prueba de Selección Universitaria

Alejandra Osses and María Leonor Varas

Introduction

Chile is a republic located on the south-west coast of South America. The country is divided into 15 regions, all dependent on a centralized government with its headquarters in Chile's capital, Santiago. According to a 2012 census, the country's population is approximately 17 million (INE, 2016).

The Chilean education system comprises 12 years of compulsory education: eight years of primary and four years of secondary education. A law passed in 2009 reorganized the length of each education cycle, assigning six years for each of them. The new organization has been operating since 2017. Children traditionally start school at the age of six or seven years old. The net enrolment rates in primary and secondary education are 95 per cent and 92 per cent, respectively (Centro de Estudios MINEDUC, 2013a).

At the end of Grade 10 (second year of secondary education) students can choose between the general and the vocational education tracks (GE and VE hereafter). GE is the pathway usually associated with university studies, while VE focuses on providing technical qualifications that serve either for further technical studies or the labour market. While the curricula in these two tracks share certain characteristics, the VE focuses more on providing qualifications in some specific technical areas, such as agriculture or industry. The enrolment in these two tracks is fairly balanced: while 55 per cent of students attend classrooms that follow the GE curriculum, 45 per cent attend classrooms where the VE curriculum is in place (Centro de Estudios MINEDUC, 2012).

In Chile, students are assessed often during their school life. For 2016, the assessment calendar issued by the Agency for the Quality of Education included learning assessments for all students enrolled in Grades 4, 6 and 10 in the subjects of reading, writing, maths, natural sciences and social

sciences. The assessment plan ensures that students are tested at least three times during their school career.

At the end of compulsory education, usually at the age of 17 or 18, students obtain their secondary education certificate provided they have achieved the grades required for this purpose. Therefore, there is no additional test at the end of schooling. Secondary education net graduation rate is estimated at around 83 per cent (Centro de Estudios MINEDUC, 2013a).

At the end of secondary education, students can either continue towards tertiary education, in a university or a vocational education centre, or enter the labour market. The net enrolment rate in tertiary education for young people aged 19 is 40 per cent, and 45 per cent for those aged 20 (Centro de Estudios MINEDUC, 2013a).

In Chile, university academic degrees are four to seven years long, whereas vocational education programmes range from two to four years. Of the 65 universities existing in the country, 36 require a university entry test: the University Selection Test (PSU – the Spanish acronym for *Prueba de Selección Universitaria*). None of the 136 VE centres in the country use the PSU assessment as an entry requirement.

The PSU is a battery of assessments in key subjects used for university selection purposes in the Integrated Admission System (SUA – the Spanish acronym for *Sistema Único de Admisión*). The SUA is under the administration of the Council of Rectors of Chilean Universities (CRUCH – the Spanish acronym for *Consejo de Rectores de las Universidades Chilenas*), a group that takes all formal and administrative decisions related to university admissions, including fundamental decisions concerning the PSU.

CRUCH

Two public and six private universities formed CRUCH in 1954. In 1981, under Pinochet's dictatorship the regional branches of the two public universities extant in Chile at the time were divided into eight public regional universities. That same year, six other public universities were created, all of which were admitted into CRUCH. In 1991, three private universities were created from the regional branches of one of the private university members of CRUCH. Finally, in 2016 two more public universities were created. Currently, 27 universities form CRUCH.

The Ministry of Education (MoE) chairs CRUCH but in practice has no power over its decisions and does not intervene in SUA's management. However, it played a significant role in the inception of the PSU by strongly pushing the idea that the test should assess the secondary education curriculum.

The CRUCH administers the SUA and all its members require the PSU for entry purposes. Since its introduction, CRUCH has delegated the responsibility of developing and administering the PSU to the Department of Educational Measurement, Assessment and Registry (DEMRE) of the University of Chile.

There are also nine additional institutions participating in SUA that are not members of CRUCH and have no decision-making powers on how the test is organized or managed. These nine universities are private and were founded mostly in the 1980s when the dictatorship government strongly promoted the proliferation of private institutions in all education levels (primary, secondary and tertiary). A few of them were created during the 1990s.

Eight of these universities entered the SUA in 2011, after an open invitation from CRUCH to all private universities in response to their criticisms for obstructing their work by delaying the publication of PSU results. Another one entered in 2016. The invitation extended by CRUCH did not include membership to the council but only the possibility of participating in the admission system. The differences in terms of public funding, which only apply for CRUCH universities, did not modify in structure due to the enlargement of the SUA.

The PSU

The PSU was introduced in the 2004 University Admission Process and came to replace the previous university admission exam, the Academic Aptitude Test (PAA – the Spanish acronym for *Prueba de Aptitud Académica*), which by that time had been in place for 37 years (Universidad de Chile, 2016). The PAA had three mandatory tests focused on general ability: verbal language, mathematics and Chilean history. (Initially there were only two mandatory tests: language and mathematics. The knowledge-oriented Chilean history and geography test was introduced as mandatory by the dictatorial government in 1984.) Another five tests were content-knowledge-oriented and served to provide more information for selection into specific academic programmes such as engineering, medicine or the sciences.

In 2000, the MoE convened a committee to re-evaluate the purpose and assessment framework of the admission tests. The authority needed to collect evidence to support the introduction of a curriculum reform introduced to secondary education in the late 1990s (Koljatic and Silva, 2006). The educational reform introduced in Chile during that decade was the first reform after the end of the dictatorship (which ended peacefully in 1990, after a referendum in 1988 and a presidential election in 1989), at a

period known as 'transition to democracy'. Because of its timing, the reform was viewed as the symbol of the returned democracy, as an opportunity to modernize the country after 25 years of social oppression and segregation and as a gateway to the new century (García Huidobro, 1999).

After an agreement reached between the Ministry and CRUCH, the PAA was abolished and replaced by PSU. The new battery of assessments comprised four tests from which individuals should take at least three. The language and communication and the mathematics tests are mandatory for every applicant. Individuals should then choose at least one of the other two assessments: the history, geography and social sciences test and the sciences test. The PSU battery of tests had a new focus – the curriculum. As defined by CRUCH and the Ministry of Education, the PSU's purpose is to assess the extent to which students have acquired the knowledge defined in the secondary education curriculum and select applicants for university (Cox, 2005; Koljatic and Silva, 2006; Pizarro, 2001).

The argument for replacing the PAA was that *because* the new assessment was curriculum based it would promote equity. The PAA was perceived as measuring something abstract – general abilities – that could be more closely related to socio-economic characteristics than to learning achievement. However, after 13 years of using the PSU, evidence shows that the gap between students coming from low and high socio-economic backgrounds has increased (Koljatic *et al.*, 2013). The main hypothesis for this problem relates to the PSU's curriculum orientation and the fact that the test assesses a large curriculum – therefore, a large amount of content.

In Chile, a highly segregated country in terms of socio-economic status, students from disadvantaged contexts tend to cluster together in the same schools – usually public (Valenzuela *et al.*, 2013). According to evidence from the Ministry of Education (Centro de Estudios MINEDUC, 2013b), public schools cover only 67 per cent and 70 per cent of the mathematics and language curricula, respectively, in Grade 12. In contrast, private schools reach 85 per cent of coverage for these two subjects. Therefore, students who graduate from some schools struggle to demonstrate their knowledge when facing a test assessing a part of a curriculum that they have not had the chance to learn.

To maintain a sense of comparability between PAA and PSU scores, the new test used the same approach to obtain final scores for each test (explained in the next section). In strict technical terms, PAA and PSU scores are not comparable because the tests measure different constructs (Coe, 2010). However, the attainment in both assessments can be compared following Newton's terminology of predictive comparability

perspective (2010); similarly graded students are thought to have a similar likelihood of future success.

All individuals wanting to apply for the universities requiring this assessment take the PSU. Therefore, the test is available for all individuals who have a secondary education certificate, regardless of their age. In the 2017 admission process, around 260,000 individuals took the PSU; of those, 72 per cent graduated from secondary school that same academic year. The remainder corresponds to individuals who graduated from secondary education in previous years.

Academic programmes have their own criteria to define a minimum average score that individuals should reach in the two mandatory tests (language and mathematics) in order to be eligible to submit an application. This minimum average score varies between programmes and universities but is not lower than 450 points – a score that could be considered as a cut score representing pass or fail.

The assessment process

The description below applies for the PSU from 2014 to the present time.

Nature of PSU assessments

The PSU is a battery of four paper and pencil tests, each with 80 multiple choice questions. Tests in language, mathematics, history, geography and social sciences and sciences assess the content of the secondary education curriculum. Table 5.1 presents test administration time for each assessment.

Table 5.1: PSU tests length

Test	Administration time
Language	2 hours 30 minutes
Mathematics	2 hours 45 minutes
History, Geography and Social Sciences	2 hours 30 minutes
Sciences	2 hours 55 minutes

The four tests are administered once a year at the end of the academic calendar (late November or early December) over two consecutive days.

After test administration, all test material returns to DEMRE for data processing and analysis. Once all response sheets are digitalized, DEMRE's analysis team evaluates the psychometric characteristics of test items and equates the different forms. The final score is estimated using the number of correct responses per person and a normalization of this distribution to

have a mean of 500 points and a standard deviation of 110 points. For the normalization, the minimum and maximum scores an individual can obtain are set to 150 and 850 points, respectively – the only difference between PSU and PAA scores is that the latter could range from 200 to 800 points.

PSU results are published on DEMRE's website 26 calendar days after the administration of the assessment. PSU results are private information; all individuals participating in the test-taking process have an account with a personal password to access their information.

Examinations

Since its introduction, CRUCH delegated the development, administration and analysis of the PSU to DEMRE, a technical department at the University of Chile. The University of Chile is the forefather of selection processes for entrance to higher education in Chile: its first oral exam (the Baccalaureate) for admission purposes was developed in 1850. Then, the test evolved into a set of different tests comprising items with open responses and essays. In 1966, when the growth in the number of applicants made the administration of tests comprising open responses and essays unfeasible, the University of Chile developed a standardized test – the PAA.

The University of Chile administered the first PAA in 1967 and, at the same time, offered the new test to the other seven universities extant in the country at the time (Universidad de Chile, 2016). The University of Chile was in charge of the entire process of defining the content, developing, administering and analysing the results of the test from 1967 to 2003. The University also processed all applications for universities within CRUCH and performed the selection of students for each institution. During this period, the university made available its technical and logistic capacity to the entire country and CRUCH members.

However, in 2003, with the change of the PAA to the PSU, DEMRE lost its rights over the test and the authority to lead and propose test changes. Currently, DEMRE develops, administers and analyses the results but has little control over modifications that can be introduced to the tests. CRUCH (which is a non-technical body) keeps control of the decisions regarding the main aspects of the tests. Thus, aspects such as the assessment framework of the PSU (i.e. curriculum-oriented), the number of questions and their format and the number of parallel forms administered each year are handled by CRUCH.

DEMRE follows strict quality assurance processes for ensuring the quality of the assessments administered each year. All items are developed and pre-tested in a sample of the target population at least a year before the

assessment. Before the pre-test at least two experts in each subject review the items.

DEMRE's teams develop each test following a specification that complies with a distribution of contents and abilities similar to that observed in the secondary education curriculum. These specifications are made public each year, around seven months before the assessment dates. Once test forms are ready for final administration, at least five content-specific experts and measurement experts provide feedback on the difficulty and pertinence of the tests.

During test administration, test-takers are given the chance to make comments about items they may consider problematic from a content point of view. In each classroom where the test is being taken, test administrators keep track of all these comments in a specific document DEMRE has designated for this purpose. Once all the test material returns to DEMRE, response sheets are machine marked and all comments are reviewed and evaluated. DEMRE also reviews an item if there are many complaints about that item in the social media.

If the result of this process reveals that an item has content problems or some kind of bias, the item is dropped from the analysis and from the calculation of scores for all individuals. If the item was only on some of the forms of the test, this elimination is considered during the equating of the different forms. Although all items go through a thorough review process, from time to time DEMRE finds problematic items that have to be eliminated from the calculation of scores.

Another reason to drop items from the calculation of scores is related to changes in their psychometric characteristics. Sometimes, items' parameters change their psychometric behaviour between the pre-test and the main administration. If an item is found to be too easy, too difficult or not having an acceptable discrimination parameter in the main administration, it is not considered in the calculation of scores. (An item is considered too easy when more than 90 per cent of the population provides a correct answer. In contrast, the item is considered too difficult when less than 10 per cent of the population provides a correct answer – 5 per cent in the case of mathematics items. The minimum coefficient accepted for the discrimination parameter is the 0.250 (biserial correlation).) This situation is unusual and affects only a few items each year. For illustration, in the 2017 admission process, only one item was dropped – from the language and communication test.

Standard setting process

Standard setting in the PSU

In terms of the development of the PSU tests, the standard setting process to ensure comparability of the assessments between years is performed at DEMRE. For this purpose DEMRE experts use specification tables that guide the test development process. The distribution of subject topics in these tables is kept stable over time, unless there is a major change in the curriculum. The last major adjustment was in 2009, when the curriculum of secondary education was modified. At that time, the specification tables were modified accordingly to maintain the alignment between the PSU and the curriculum.

To give meaning to PSU scores, there is no formal standard setting process either. Universities use their own criteria and experience to make all decisions regarding this matter. The fact that universities do not select individuals with an average score in mathematics and language tests lower than 450 points has no formal foundation. However, we could say that according to their experience 450 points is the minimum average score acceptable in the two mandatory tests for admitting students – an argument that could be interpreted as a standard setting process based on experience.

The results of each PSU administration are cohort-referenced. Scores are not comparable over time because every year individuals taking the test reconstruct the items in social media. With these reconstructions, most of the items are made public and it is impossible for DEMRE to repeat some of these over time in order to ensure valid comparisons of scores. The relative ranking of individuals can be compared over time using the percentile distribution in the scale of scores.

Considering that the mean score is 500 and the maximum score reaches 850 points, scores over 650 or 700 points in the tests are considered good enough for some academic programmes but insufficient for others. However, these appraisals are subjective and depend on how selective the academic programme is that the individual wants to pursue, the number of places offered and the number of applicants.

Standard setting in the university application process

University admission does not depend solely on PSU scores. Until 2012, selection depended on two factors: PSU scores and secondary education GPA. Since 2012, CRUCH includes a third selection factor to be taken into account in SUA – the GPA ranking score, which represents the relative position of the applicant in all the education contexts (schools) in which he or she pursues secondary studies. Within universities, each academic

programme awards different weights to these three factors to calculate an application score. The combination of PSU scores should represent at least 50 per cent of the application score.

Universities not only define a minimum average score applicants should obtain in the language and mathematics tests in order to be able to submit an application to their institution; they also set a minimum application score. This score is – according to their experience – the minimum standard acceptable to pursue studies in their programmes.

In general, we could say that institutions assign weights to the selection factors and define these minimum scores according to the type of students they want to attract. For example, universities targeting disadvantaged students and concerned about implementing affirmative action would give less importance to PSU results and more weight to the ranking score. Selective universities, concerned with maintaining high academic standards, may prefer to assign more importance to PSU and require higher minimum application scores.

Table 5.2: Comparison of weights for the medicine programme application score between four universities, 2017 admission process

		University			
		1	**2**	**3**	**4**
Weight of selection factors (in %)	GPA	10	20	15	10
	Ranking	30	20	25	40
	L&C	10	15	15	10
	Maths	25	20	35	20
	HG&SSc	0	0	0	0
	Sciences	25	25	10	20
Minimum average between L&C and Maths tests requested		450	475	475	475
Minimum application score requested		600	600	500	600
Number of individuals selected		172	93	115	61
Maximum application score selected in 2017		836.00	828.80	834.80	795.20
Lower application score selected in 2017		786.20	796.75	767.90	782.10

Rank: ranking score; L&C: Language and Communication; HG&SSc: History, Geography and Social Sciences

Table 5.2 presents a real example of the weights assigned to the same academic programme in four different universities for the 2017 admission process (CRUCH, 2016). The example is a very selective programme – medicine. As we can observe, three of these universities assign 60 per cent of weight to the combination of PSU scores. University number 4 assigns the minimum allowed for PSU (50 per cent) and a significant percentage to the ranking score, looking to attract students who performed well compared to their schoolmates.

Table 5.3: Additional admission criteria for teaching academic programmes

Year	PSU criteria		GPA Ranking		Other criteria
2017	Average score of at least 500 points in Language and Mathematics tests	or	Being in the top 30% of students	or	• Graduate from an admission programme certified by the MoE
2020	Average score of at least 525 points in Language and Mathematics tests	or	Being in the top 20% of students	or	• Being in the top 40% of students in the GPA ranking AND obtain at least an average score of 500 points in Language and Mathematics tests • Graduate from an admission programme certified by the MoE
2023	Average score of at least 550 points in Language and Mathematics tests	or	Being in the top 10% of students	or	• Being in the top 30% of students in the GPA ranking AND obtain at least an average score of 500 points in Language and Mathematics tests • Graduate from an admission programme certified by the MoE

In 2016, the Ministry of Education introduced additional criteria for applicants to teach academic programmes in all universities in the country. These changes form part of a wider reform that seeks to improve the social

appraisal of the teaching profession for future generations. The additional criteria for admission to these academic programmes are described in Table 5.3 and are intended to promote the admission of high achieving students.

Applicants can apply to a maximum of ten academic programmes, ranking this selection according to their preferences. For each academic programme, individuals are ranked on their application score. Their selection depends on the position in this rank and the number of places offered in the academic programme.

The selection process is performed using a 'stable matching algorithm' that ensures an optimal assignment for both applicants and academic institutions. David Gale and Lloyd Shapley first published this allocation mechanism in an abstract theoretical setting in the early 1960s (Gale and Shapley, 1962). In 2012, Alvin Roth and Lloyd Shapley were awarded the Nobel Prize in Economics for the development of this theory and its relevant applications in this field (Royal Swedish Academy of Sciences, 2012). It is very remarkable that the Chilean university selection process is one of the oldest, and most widely adopted, applications of this optimal procedure.

Political, public controversies and debates with the PSU

The origin of the public controversies surrounding the PSU can be situated, mainly, in the different modifications introduced to SUA over time. In the last decade, the admission system has been in need of adjustment due to the growth experienced in tertiary education access. In Chile, the number and variety of tertiary education institutions have increased significantly, and the lack of regulation of the sector has become a national issue. Only the most regulated portion of these institutions – those that are academically oriented – use the PSU.

The education reform currently under discussion in parliament ties the public funding of tertiary education to the regulation of admission procedures. The bill sent by the government proposes the use of an admission system for all tertiary education institutions that includes a variety of assessment instruments. These instruments should be designed according to the diversity of the institutions and programmes considered in this system – not just universities but also vocational education institutions.

In this context, two urgent needs arise: the development of new instruments and the adjustment of the current admission tests. The ongoing discussion, which prefigures a new system with new tests that surpass the current criticisms, results in a loss of value of the PSU. This scenario is very delicate because new tests cannot be ready for use in less than four years.

The new instruments should assess the knowledge, skills and content-specific competencies of applicants coming from a wider variety of educational contexts. They should also meet the selection requirements of a wider variety of institutions, ranging from academically oriented universities to vocational schools.

In the meantime, improvements to the PSU are urgently needed. In this debate, considerations should be given not only to the characteristics of the current tests for reproducing the inequity of an already highly segregated schooling system but also to those who have the power of modifying these characteristics.

As we mentioned earlier, the reproduction of social inequities increased when the PSU was introduced, presumably due to the curriculum-oriented focus of the test. This behaviour contradicts the promises made by the authorities and the promoters of the PSU, who created the expectation that the new tests would promote equity, allowing the inclusion of a sector traditionally excluded from higher education: those coming from disadvantaged contexts (Koljatic and Silva, 2006; Koljatic *et al.*, 2013).

The introduction of the PSU also meant giving important powers into the hands of CRUCH. This council, supported by the Chilean MoE, exerted a fierce defence of its decisions, disregarding any criticism or questioning. CRUCH and the government interpreted criticisms of the new tests as political attacks. Since the return to democracy in 1990, Chile invested a significant amount of financial resources (mostly coming from international loans) in developing and implementing a curricular reform. According to the authorities, the success of this reform should be reflected in an assessment such as the PSU. Therefore, the increasing socio-economic gaps in PSU results were a problem that the MoE and CRUCH chose to ignore. Due to their lack of reaction, the resistance grew and those against the new test are accumulating reliable research evidence to support their critical position (Koljatic and Silva, 2006; Koljatic *et al.*, 2013; Larrañaga *et al.*, 2014).

While CRUCH presented evidence supporting the claim of stability or a small increase in the achievement gaps between different socio-economic groups, detractors of the PSU produced other findings. Differences between these two groups were not only at the methodological level. The PSU curriculum alignment reflects only a part of the curriculum – that of the GE track – ignoring the existence of an important part of the school population – the VE sector – whose curriculum does not offer the opportunity of learning the content assessed by the PSU. Thus, an important part of the studies developed by PSU supporters simply did not consider the VE population

when studying socio-economic gaps in achievement (CTA–CRUCH, 2009; Koljatic and Silva, 2006, 2013; Larrañaga *et al.*, 2014).

CRUCH official studies estimated differences in achievement between private and public schools, adjusting differences by individual socio-economic status (measured by the monthly family income declared when registering to take the PSU). They concluded that achievement gaps did not increase over time (see Figure 5.1, Graph 1; CTA–CRUCH, 2009). However, that result does not consider the population who graduated from vocational schools, a fact that was not noted either in the report that made these results public. When including VE population in the analysis and using the same methodology of the CRUCH study, Silva and her colleagues (2016) demonstrate the increasing gap and the harm for VE students. In Figure 5.1, Graph 2 reveals the evidence masked by Graph 1. Seeing both graphs together, we can conclude that CRUCH's claim of gaps being stable over time is not sustained.

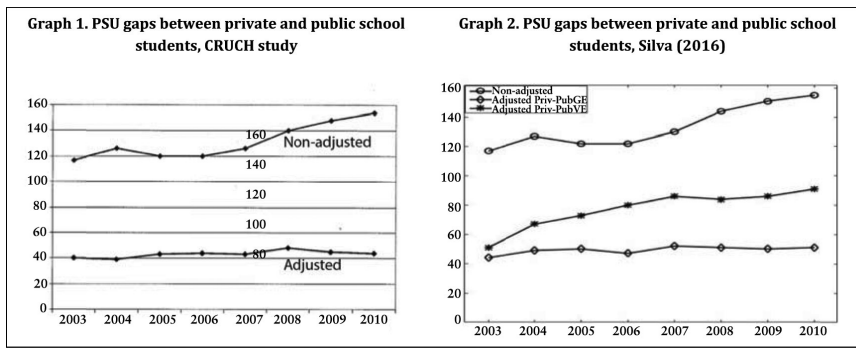


Figure 5.1: PSU gaps between private and public school students, CRUCH study (Graph 1); PSU gaps between private and public school students, Silva (2016) (Graph 2)

Priv: private school; Pub: public school; GE: General Education track; VE: Vocational Education track. Adjusted scores refers to those obtained after controlling PSU results by student socio-economic level (measured by the monthly family income reported at the time of registering to take PSU).

Other criticisms of the PSU relate to the misuse of results. The MoE uses the PSU results to assign student scholarships and provide extra funding to those universities enrolling students with higher scores. Each year, the Ministry provides scholarships covering part of the annual fee to the students achieving the top score in any of the PSU tests – 850 points – and to students reaching the higher average score in the language and mathematics tests (MoE, n.d.). All graduates from schools receiving public funding and belonging to the 80 per cent of population with lower family income are eligible for the scholarships, if they reach the aforementioned score.

Until 2016, education institutions also received extra funding from the MoE for enrolling students among the 27,500 best scores in the PSU (known as Indirect Public Contribution). Until 2015, the available funding was around US$38,500,000 to be shared by universities or VE institutions. In 2016, this funding was reduced by 50 per cent due to the start of the free-of-charge education policy for the 50 per cent of the population with lower family income and enrolled in universities. The reduction of the Indirect Public Contribution placed the MoE in a difficult situation because it meant a significant reduction of funding for some institutions. In fact, one initiated court actions.

The use of PSU scores to assign funding to institutions or scholarships is problematic, because there is no evidence supporting the validity of using results for this purpose.

Thirteen years after introducing the PSU, and two international audits that endorsed the main criticisms of the tests (Educational Testing Service, 2005; Pearson, 2013), both the system that created the PSU and the foundations of the tests have been weakened. In the definition of the twofold purpose of the PSU – provide evidence to support the curriculum reform and select applicants to university – there was a need to justify the introduction of a curricular reform, a point of inflection for the education sector after 17 years of dictatorship. This need of the government required adherence that stigmatized criticisms.

Chile's current needs are of a different nature. In addition to the reform of tertiary education (which is being discussed in parliament), in the last two years the government has initiated a free access programme to universities and other tertiary education institutions for students coming from disadvantaged socio-economic backgrounds. The public funding for these students is tied to the use of objective and transparent selection processes. Therefore, even before the development of the new Common Admission System, considered in the bill proposed by the government, there is a current need to broaden the variety of tests and instruments. These need to be appropriate to meet the requirements of fair admission processes and target a more complex and diverse system and population.

Although the parliamentary debate seems to have a long way to go before any consensus is reached, the need for changes to the PSU tests is now urgent. It seems that we have consensus about the need to reduce the curricular content assessed by the PSU, in order to ensure that all applicants had been exposed to them.

Recent research conducted by DEMRE (Silva *et al.*, 2016), reproduced a previous predictive validity study requested by CRUCH (Grau, 2016). In

the study, items measuring Grade 11 and 12 curriculum content (where GE and VE differ) were dropped from the analysis in the mathematics PSU test. The predictive validity study performed without these items shows that there is no loss of reliability or predictive validity with the reduced test, but a lower difficulty and a small reduction of the socio-economic gap in the results. In ongoing studies with the other PSU tests that DEMRE is currently developing, experts recommend a reduction of assessed curricular content and focus on relevant educational achievements instead of the current detailed curricular coverage (DEMRE, in development). These are important arguments to drive some urgent improvements of the PSU, related to fairness.

However, today's problem is not only about fairness and students who are being assessed in a curriculum they have not had the opportunity to learn; the criticisms have gone beyond that point and focus on the lack of validity of using PSU results for university admission.

For instance, an additional subgroup for which these results are not valid has been defined directly by the Ministry. The Programme of Support and Effective Access to Higher Education (*Programmea de Acompañamiento y Acceso Efectivo a la Educación Superior* in Spanish), an initiative of the MoE, defined different admission criteria for students participating in the programme. These criteria deliberately leave PSU results out of the selection algorithm, acknowledging a validity problem. Yet, the current criteria fail to generate unique assignments of students to academic programmes. In case of a 'tie' (two students with exactly the same values in the selection factors and applying to the same programme), the selection is solved by lottery instead of using academic criteria such as those provided by a test such as the PSU.

Universities also claim that PSU tests do not provide information on what students know and are able to do, arguing that students' starting academic level is below what is expected. By focusing only on content knowledge, tests are not assessing individual skills on how to apply that knowledge in a university context. Since universities are not receiving information about the academic skills of their newly enrolled students, university curriculum and instruction is disjointed from students' starting academic level. If the current tests do not provide the information required by universities as part of an admission system, there is also a lack of validity regarding their use for selection purposes.

In the context of a new four-year-long R&D project recently awarded to DEMRE, a battery of new instruments will be developed. The aim of these instruments will be the assessment of content specific competencies in

mathematics, language and science and general non-academic skills that are highly predictive of academic success. The MoE supports and participates in this project and is interested in the alignment of these new instruments with the fundamental competencies and skills promoted in the school curriculum and the non-academic skills promoted by the official school accountability policy.

The R&D project is also interested in exploring the use of a written exam and the use of some constructed-response items. Currently, PSU uses multiple choice items because of the ease of marking and the limited timeframe set by CRUCH between test administration and publication of results (26 calendar days). However, there is an ongoing debate about the limitation of this kind of item to assess some crucial competencies and abilities needed in higher education, such as argumentation (McCurry and Orpwood, 2012; Soland *et al*., 2013).

These initial definitions and agreements for developing new instruments and new standards are, of course, preliminary. A complete research programme has to provide solid evidence about the behaviour of socio-economic gaps and the expected benefits of the innovations. There is an explicit design of the procedures to define, review, improve and update these standards. These new procedures consider the participation of a wide spectrum of stakeholders and the use of empirical evidence to support all decisions. From DEMRE we have promoted technical, social and political discussions that allow us to progress in this direction. We have sound reasons to be optimistic and we believe that we will see changes in the short-term future.

References

Centro de Estudios MINEDUC (2013a) *Chile en el panorama educacional internacional OCDE: Avances y desafíos. Evidencia* series no. 18. Online. https://goo.gl/MzdQvU (accessed 8 June 2018).

Centro de estudios MINEDUC (2013b) *Implementación del currículum de educación media en Chile. Evidencia* series no. 21. Online. https://goo.gl/kDcaMz (accessed 8 June 2018).

Coe, R. (2010) 'Understanding comparability of examination standards'. *Research Papers in Education*, 25 (3), 271–84.

Cox, C. (ed.) (2005) *Políticas educacionales en el cambio de siglo. La Reforma del sistema escolar de Chile*. Santiago de Chile: Editorial Universitaria.

CRUCH (Consejo de rectores de las universidades chilenas) (2016) *Oferta definitiva de carreras, vacantes y ponderaciones proceso 2017*. Santiago: DEMRE. Online. https://goo.gl/2dkkBz (accessed 8 June 2018).

CTA (Consejo técnico asesor)–CRUCH (2009) *Presentación en reunión del consejo de rectores*. Santiago.

DEMRE (2016) *Evidencia Preliminar para la Reducción de Contenidos de la PSU de Matemáticas*. Online. http://psu.demre.cl/estadisticas/informe-tecnico (accessed 27 July 2018).

DEMRE (in development) *Revisión de los marcos teóricos de la PSU*.

Educational Testing Service (2005) *Evaluación externa de las pruebas de selección universitaria (PSU)*. Online. https://ciperchile.cl/wp-content/uploads/informe-ets.pdf (accessed 27 July 2018).

Gale, D. and Shapley, L.S. (1962) 'College admissions and the stability of marriage'. *American Mathematical Monthly*, 69 (1), 9–15.

García Huidobro, J.E. (ed.) (1999) *La reforma educacional chilena*. Madrid: Editorial Popular.

Grau, M. (2016) *Estudio acerca de la validez predictiva del ranking de notas*. Sistema único de admisión – CRUCH.

INE (Instituto nacional de estadísticas de Chile) (2016) *Demográficas y vitales*. Online. www.ine.cl/canales/chile_estadistico/familias/demograficas_vitales.php (accessed 12 July 2018).

Koljatic, M. and Silva, M. (2006) 'Equity issues associated with the change of college admission tests in Chile'. *Equal Opportunities International*, 25 (7), 544–61.

Koljatic, M. and Silva, M. (2013) 'Opening a side-gate: Engaging the excluded in Chilean higher education through test-blind admission'. *Studies in Higher Education*, 38 (10), 1427–41.

Koljatic, M., Silva, M. and Cofré, R. (2013) 'Achievement versus aptitude in college admissions: A cautionary note based on evidence from Chile'. *International Journal of Educational Development*, 33 (1), 106–15.

Larrañaga, O., Cabezas, G. and Dussaillant, F. (2014) 'Trayectorias educacionales e inserción laboral en la enseñanza media técnico profesional'. *Estudios públicos*, 134, 7–58. Online. https://goo.gl/82F1tF (accessed 8 June 2018).

McCurry, D. and Orpwood, G. (2012) 'Assessing skills for success in tertiary education'. Australian Council for Educational Research. Online. https://works.bepress.com/doug_mccurry/36/ (accessed 8 June 2018).

MoE (Ministerio de Educación) (2012) *Educación Técnica Profesional en Chile: Antecedentes y Claves de Diagnóstico*. Santiago: Centro de Estudios del Ministerio de Educación de Chile.

MoE (Ministerio de educación) (n.d.) *Beca puntaje psu (bpsu)*. Online. http://portal.beneficiosestudiantiles.cl/sites/default/files/detalle_bpsu2018.pdf

Newton, P.E. (2010) 'Contrasting conceptions of comparability'. *Research Papers in Education*, 25 (3), 285–92.

Pearson (2013) *Final report: Evaluation of the Chile PSU*. Online. http://educacion2020.cl/wp-content/uploads/2013/01/201301311057540.chile_psu-finalreport.pdf (accessed 12 July 2018).

Pizarro S.R., (2001) 'Nueva P.A.A. Chilena: Algunas Consideraciones Políticas, Teóricas, Técnicas y Funcionales', *Revista de Psicología de la Universidad de Chile*, X (1). Online. http://repositorio.uchile.cl/bitstream/handle/2250/122264/nueva-PAA-chilena-algunas-consideraciones-politicas-teoricas-tecnicas-y-funcionales.pdf?sequence=1 (accessed 27 July 2018).

Royal Swedish Academy of Sciences (2012) 'Stable matching: Theory, evidence, and practical design'. Online. https://goo.gl/4zDWW3 (accessed 8 June 2018).

Sistema Único de Admisión (2017) *Estudio acerca de la validez predictiva del ranking de notas*. Consejo de Rectores de las Universidades Chilenas (CRUCH). Online. http://sistemadeadmision.consejoderectores.cl/public/pdf/publicaciones/Libro_Ranking_Notas(web)_baja.pdf (accessed 27 July 2018).

Soland, J., Hamilton, L.S. and Stecher, B.M. (2013) *Measuring 21st Century Competencies: Guidance for educators*. Rand Corporation. Online. http://asiasociety.org/files/gcen-measuring21cskills.pdf (accessed 8 June 2018).

Universidad de Chile (2016) Historia del examen de admisión. Online. https://goo.gl/4vxvuC (accessed 8 June 2018).

Valenzuela, J.P., Bellei, C. and de Los Ríos, D. (2014) 'Socioeconomic school segregation in a market-oriented educational system: The case of Chile'. *Journal of Education Policy*, 29 (2), 217–41.

The consequential dimension of validity in the Chilean University Entry Test

María Teresa Flórez Petour

Chapter 5 by Alejandra Osses and María Leonor Varas constitutes a good summary of the main features of the Chilean university entry test (PSU). The authors also raise important issues around validity in connection with the purposes, constructs, predictability, potential socio-economic bias and lack of meaningfulness of items in these tests. Public controversies around the PSU, however, are only addressed to the extent that they involve the actors that have a direct influence in decision-making processes with regard to the test. There are other actors who are importantly affected by the PSU, and their consideration would also be relevant to the future development of a more valid selection system, especially since the introduction of consequences as an important dimension in validity studies (Messick, 1979; AERA *et al*., 2014).

Research in relation to the effects of the PSU in the work of teachers and schools is only incipient. A recent qualitative study by Gazmuri (2017) explores the effects of the test in teaching and learning in the area of history. Initial findings reveal a significant impact on teachers' work: classrooms that are influenced by the test work with a more limited and basic range of skills, mainly memorization and application; activities are therefore less ambitious and focused on content-based materials aimed at training students to the test, to the detriment of group work or more challenging and complex tasks (Gazmuri, 2017). Similarly, Flórez Petour (2014) highlights how teachers experience role conflict in connection to this and other national testing systems, in terms of feelings of tension and contradiction between the need for covering all the content that constitutes the syllabus of the university entry test, and the relevance they attribute to other skills and attitudes that they believe students should develop for their future lives. In the highly competitive and market-oriented Chilean education system, schools that hold among their promises to parents that students will enter higher education are the ones that experience these consequences more strongly. At an individual level, competition and social differentiation of results are also connected to the possibilities of families who have access

to *Preuniversitarios*, a whole business area that has emerged in connection with the preparation for these tests in private training centres.

Among the changes that are proposed for the future of the PSU, thought should be given to its consequences and underlying pedagogies, including the development of further research around the effects of the test. Authenticity should be a central concern, in terms of designing tasks that are able to motivate the development of complex skills through meaningful, real-life, problem-centred activities that are not easy to train in a mechanistic fashion. In addition to this, the wider spectrum of stakeholders to be considered in the design of the new tests that Osses and Varas refer to should undoubtedly include teachers and students. These considerations would solve not only issues around the predictability of these tests in connection with future academic performance in higher education but would also allow schools to focus on what they deem relevant for the future lives of young people, independent of whether their plans include the aspiration to university studies.

References

AERA (American Educational Research Association), APA (American Psychological Association) and NCME (National Council on Measurement in Education) (2014) *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Flórez Petour, M.T. (2014) 'Assessment reform in Chile: A contested discursive space'. Doctoral thesis, University of Oxford. Online. https://goo.gl/QA76rr (accessed 9 June 2018).

Gazmuri, R. (2017). 'Estudio cualitativo de las paradojas curriculares de la PSU de Historia'. Preliminary findings of a FONDECYT project on curricular paradoxes of the PSU in History, Geography and Social Sciences. Video, 28 December. Online. https://goo.gl/FAoY7U (accessed 9 June 2018).

Messick, S. (1979) 'Test validity and the ethics of assessment'. *ETS Research Report Series*, 1979: i–43. Online. https://goo.gl/CNLWKP (accessed 9 June 2018).

Setting standards in the Chilean university entry test

Francisco Javier Gil Llambías

This work provides an excellent historical description of the changes in the system of access to Chilean universities in the past decades. It is very positive that the authors expose the strengths and weaknesses of the Integrated Admission System (SUA). The main strength of the SUA is that students from the most remote cities and towns can apply to one of the 1,897 courses offered by 39 universities located thousands of kilometres away, having taken the same battery of tests. It is also a strength that the Department of Evaluation, Measurement and Educational Registration (DEMRE) currently has a Development and Analysis Unit with the capacity to propose improvements to evaluation instruments, such as those summarized in this chapter. Likewise, it is a strength that the DEMRE has an internal control system that makes corruption cases virtually impossible.

However, the system contains serious weaknesses, some of which can be glimpsed in this chapter:

1. In the 1990s the Ministry of Education (MoE) decided to create a single test with two different purposes: (a) to select applicants for university vacancies and (b) to assess the extent to which students had acquired the knowledge defined in the secondary education curriculum. This decision increased the inequity of the system because public and private schools cover less than 70 per cent and near to 85 per cent of the official curriculum, respectively, in Grade 12.
2. From 1981 until 2016 education institutions received annually extra funding of around US$38,500,000 from the MoE for enrolling students among the 27,500 best scorers in the PAA/PSU (known as Indirect Public Contribution – AFI). With the objective of optimizing the income of AFI, universities raised the weighting of the PAA, causing the most harm to students on applied programmes in state schools, where the poorest students study. Moreover, state scholarships are assigned only to students who surpass certain minimums of PSU. For example, the scholarship to study education requires a minimum score of 600 PSU points, which 99 per cent of the students who graduated in 2016 from technical high schools – where the poorest students study – did not

reach. Thus, the PSU is used to prevent the poorest students studying pedagogy.

Fortunately, since 2012, the CRUCH (Council of Rectors of Chilean Universities) included a third selection factor to be taken into account in SUA: the GPA ranking score. This new factor lowered the weight of the PSU from 71 per cent to 60 per cent, between 2012 and 2017. The gaps between the average of PSU scores and the GPA ranking of the 2016 cohort of students who graduated from private and state schools were 137 and 22 points, respectively. On the other hand, studies have shown that the GPA ranking better predicts academic behaviour than the PSU, especially after the first year of university studies. Since 2014, 29 CRUCH universities have offered special places to students with a GPA ranking higher than 702 (the top 15 per cent of students) exempted from the PSU score, who graduated in the 456 poorest schools of Chile.

The GPA ranking and the new tests developed by the Development and Analysis Unit of DEMRE are reasons to be optimistic, and we believe that we will see changes in the short-term future.

Chapter 6

Standard setting in England: A levels

Rachel Taylor and Dennis Opposs

Introduction

England is one of the four nations of the United Kingdom (UK). It is located to the west of continental Europe and comprises the central and southern part of the island of Great Britain. With a population of 55.3 million, it has 84.2 per cent of the population of the UK (ONS, 2016).

Students in England usually attend school from age 4. At age 16, almost all pupils take General Certificate of Secondary Education (GCSE) exams, each based on a different subject. A typical student takes about nine GCSEs including English, mathematics and a combination of other subjects. Since 2015, all 17- and 18-year-olds have to be in some form of education or training, with the majority choosing to continue in education (DfE, 2017). For these students, a wide variety of courses is available, from academic subjects through to vocational courses equipping students directly for the world of work. Between the ages of 16 and 18, most students in full time education study Advanced Level General Certificates of Education (A levels). By age 19, around 40 per cent of young people in England enter higher education at universities and colleges (UCAS, 2017). In 2017, 391,370 students from England were accepted onto higher education courses in the UK (UCAS, 2017).

Applications to higher education across the UK, including English universities, are centralized and managed by UCAS. To secure a place on an undergraduate course, students can use a range of qualifications (including A levels). Each university, and courses within them, typically have their own entry requirements in terms of the qualifications and grades required. Other factors, such as the quality of an applicant's personal statement and references may also be taken into account when applicants are considered for course places. Single-subject courses of full-time study are typically three years for an honours degree, although two-year foundation degrees are also available.

Education policy direction for England is set by the Department for Education (DfE) supported by a number of government bodies. Ofqual regulates the assessments for qualifications available within state-funded English schools (GCSEs and A levels, for example) and the various exam boards that develop and provide them. Ofqual also accredits new GCSE and A level qualifications that the exam boards develop. Each of these qualifications has its own procedures for maintaining (linking) standards over time.

A levels

A levels were introduced in 1951 and are generally taken by 16- to 18-year-olds in schools and colleges. However, they are available to anyone who would like to gain a qualification in a subject that they are interested in. A levels are the principal pre-university qualification in England. In many cases, students need to have gained at least five GCSEs at grades A*–C in order to study A levels. Although entry requirements vary across schools and colleges, almost all education providers also stipulate requirements for students to have achieved a certain GCSE grade in a particular subject before continuing to study that subject at A level.

A levels are available in over 45 subjects and can be studied alongside other qualifications. There is no compulsory subject element and A levels can be taken in any combination desired to reflect the interests (or intended progression) of the student. In 2016, there were 743,986 A level subject entries (DfE, 2016).

An A level is typically taken over a two-year period. From 2017, new A level qualifications were awarded following reforms instigated by the government that was elected in 2010. The reforms are phased in over three years for different subjects. In these reformed qualifications, students can choose to sit a standalone advanced subsidiary (AS) qualification after the first year of study, and/or sit the full A level after the second year of study. Assessments are only available in the summer exam series and all assessment is carried out at the end of the course. Students typically take three or four subjects at A level, although the effect of the introduction of the reformed qualifications on students' entry approaches is not known at the time of writing.

The A level is a graded qualification. The passing grades are A*, A, B, C, D and E. The A* grade was introduced in 2010 in response to concerns that there was insufficient differentiation at the top of the grade range (DfES, 2005). Those not achieving the lowest pass grade of E are reported as unclassified (U). In the summer 2017 exam series, the pass rate (grade E

and above) for all A level qualifications in England was 97.9 per cent, with 26.3 per cent of entries being awarded a grade A or above (JCQ, 2017).

A levels are available from four exam boards in England: the Assessment and Qualifications Alliance (AQA), Edexcel (a part of the Pearson group), Oxford, Cambridge and RSA Examinations (OCR, part of Cambridge Assessment) and the Welsh Joint Education Committee (WJEC/Eduqas). All four exam boards offer A levels in most subjects, and schools, colleges and individuals can choose between them. This means that Ofqual, the exams regulator, has a key role in ensuring that there is comparability of grade standards between the exam boards in each subject. For example, a grade B in chemistry awarded by AQA has to be of a comparable standard to a grade B in chemistry awarded by Edexcel or OCR or WJEC, since grades are used to compare individuals applying for higher education and employment opportunities.

AS and A levels have clear guidelines setting out how the qualification should be set up, what students need to learn and the skills they need to develop. The exam boards must make sure that the qualification that they offer in each subject meets these criteria before it can be offered to schools. A 'syllabus' provides the course content and details of assessments, including marking criteria that are used for any school-based assessment (coursework).

The assessment process

The descriptions below refer to typical A level practice in the years leading up to 2016. The majority of these descriptions are also relevant to the reformed qualifications, first assessed in 2017.

Nature of assessments

Each A level qualification contains between four and six units of assessment. Units are assessed either by an exam set by the exam board or on the basis of work completed over a longer period of time (school-based assessment typically referred to as 'coursework'). Examined units are up to three hours long, typically contain one (paper-based) written exam and are available once each year (between May and June). Coursework is typically assessed in the school or college.

The permissible balance of exams and coursework is prescribed for each subject. A typical A level will have approximately 30 per cent of the total score allocated to coursework, but some have none at all and more applied subjects might have up to 67 per cent of the total score composed of coursework. Students achieve marks and grades for each unit, which are

then combined to give an overall qualification grade. Students are permitted to resit A levels but generally must resit all units for a given qualification in a subsequent exam series (usually the following summer; in general, a student's coursework mark may be carried forward from a previous exam series).

Exams

The format of examined units varies by subject but typically includes a combination of multiple choice, short response and longer essay-style questions (see Annex 1). The question papers and mark schemes for examined units are drafted by a principal examiner who is responsible for that unit, often up to two years before the exam is sat. The draft assessment materials undergo several stages of review to ensure coverage of the content, comparability with previous papers, clarity and so on. This typically involves a review by other senior examiners under the guidance of the chair of examiners for that subject.

After any revisions, the paper goes to an assessor or scrutineer who checks that it is fair to candidates and can, for example, be completed in the time allowed. The chair of examiners signs off the final version. New papers are produced each summer and questions are not pre-tested. Although each year's set of papers is produced on the basis that it will be of the same demand as in previous years, in practice, it is very difficult to produce exactly the same level of demand and so cut scores are determined each summer to maintain standards – a process also known as awarding.

School-based assessment (coursework)

Coursework tasks (such as presentations, essays and portfolios) are designed to assess students' performances against assessment criteria set out by the exam board in the syllabus for the subject. While all students taking a specific subject with a specific exam board are assessed against the same criteria, there is often scope for the schools or students to set the topics that the tasks are based on. In these cases, schools can be required to obtain prior approval from the exam board on the topic chosen. Coursework tasks typically do not change from year to year.

Marking students' work

Marking completed exam papers

For examined units, completed papers are marked on paper or electronically (on-screen) by examiners. Examiners are recruited based on their subject expertise, are often teachers and are each given an allocation of papers to mark. Before marking of students' exam papers commences, the principal examiner for each unit convenes a standardization meeting with markers

(this may be either face-to-face or virtual) to ensure that they all interpret the mark scheme in the same way. At this meeting, the markers also score a number of common scripts and review their marks to confirm that they are working consistently. Before they can start marking their allocation, examiners have to demonstrate that they are applying the required standard correctly.

During the marking period, which is usually about three weeks for each paper, each examiner's work is quality checked by their respective exam board to ensure that their marking is consistent and to the required standard. The types of check vary depending on whether scripts are marked on paper or on-screen, as well as whether they are marked by question or as a whole paper.

Where marking is conducted on-screen (as most now is), checking is carried out by including 'seeded' items randomly through the marking or by double-marking. 'Seeds' are responses that senior examiners have previously reviewed and for which they have agreed a mark. Examiners are not aware which items are 'seeds' and can be stopped from marking if they do not mark the 'seed' to the agreed standard. For longer-response items, some exam boards use double-marking, where a sample of each examiner's allocation is marked by another examiner. If the marks of the two examiners are not within an agreed tolerance, a senior examiner adjudicates. Examiners who are not marking in line with the required standard can be stopped from marking. Examiners who are stopped from marking are unable to mark any further responses until they have spoken to a more senior examiner. Where there are lingering concerns over an examiner's marking, they can be stopped from marking altogether.

Where scripts are marked on paper, examiners send samples of their marking to a more senior examiner for checking. If an examiner is not marking to the required standard, they are not allowed to continue and their scripts are allocated to a different examiner.

Marking school-based assessment (coursework)

For coursework units, teachers within a school or college assess their students' work against the assessment criteria provided by the exam boards. Their marks are submitted to the exam board and a sample of marked students' work is subject to a moderation process to check that the marks have been awarded in line with the agreed standard.

Moderators are employed by the exam boards and undergo standardization to ensure that they have a common understanding of the mark scheme. If the original marks from the school or college are consistent

with those of the moderator (within an agreed tolerance) then the original marks are accepted. If the original marking is outside an agreed tolerance, then the moderator marks a further sample and the marks are analysed to determine whether the marks from the school or college need to be adjusted.

Moderators' work is checked at regular intervals during the moderation process by senior moderators to ensure that their judgements are consistent and in line with the agreed standard.

Standard setting process

Determining grades

When the majority of exam scripts in a subject have been marked, an awarding meeting (standards maintaining meeting) is convened to recommend grade boundary marks (cut scores) for grades A and E in each A level subject. Awarding committees are chaired by a senior examiner who has overall responsibility for standards in each subject. The committees generally also include a chief examiner, principal examiners (responsible for examined units), principal moderators (responsible for coursework units) and exam board technical experts. The awarding period typically lasts around four to five weeks from the end of June to the beginning of August. Awarding meetings were traditionally conducted face-to-face and lasted one to two days. More recently, most exam boards have developed online systems for undertaking parts of the awarding process remotely.

The basic principle behind the standards maintaining process for A levels is to retain from year to year the level of performance at a grade boundary mark. As stated above, although examination papers are produced on the basis that they are of the same demand as previous years, in practice it is very difficult to produce exactly the same level of demand so maintaining this standard is challenging. To help them meet the principle, exam boards draw on a variety of sources of evidence when setting grade boundaries, using both statistical and judgemental techniques.

The main statistical evidence takes the form of prior attainment-based predictions at the cohort level. Prior attainment-based predictions map the relationship between prior attainment (mean GCSE score) and A level outcomes for the cohort of students taking each subject in a reference year, and use this relationship to predict the outcomes for the current cohort of students based on their prior attainment. As such, if the prior attainment of the cohort remains similar, then the outcomes would be expected to be similar.

The predictions for each A level cohort are typically generated for 18-year-old students and are therefore based on the GCSE results that the

students obtained two years earlier when they were 16 years old. These predictions guide the awards, helping to determine the grade boundary marks, which are then applied to all students. Prior attainment-based predictions have been used by the exam boards to guide the maintenance of standards for the last couple of decades, though not necessarily in a consistent manner. This changed following the introduction of the comparable outcomes approach by Ofqual in 2010 for A levels and 2011 for GCSEs (Ofqual, 2015). The basic premise of this approach is that if the nature of the cohort sitting a qualification each year does not change, then the outcomes should not change either.

In addition to the statistical predictions, judgemental evidence is used in setting grade boundaries. This includes expert scrutiny of students' scripts by awarding committee members, reports from the principal examiners and moderators and descriptions of the expected performance at each key grade. The main source of judgemental evidence is script scrutiny. Examiners are presented with exam scripts in a range of marks (typically three to five marks) as guided by the statistical evidence, and must independently decide whether each exam script in the range is worthy of the grade under consideration. In doing this, examiners are able to refer to archive scripts on the grade boundary marks from previous years and statistical evidence showing the performance of individual questions on each exam paper. The examiners' judgements are recorded on what is known as a 'tick chart', as shown in Table 6.1. A tick means that a committee member thinks that the work is worthy of the higher grade of the boundary pair (e.g. A/B), a cross means that they do not and a question mark means that they have some doubts. Based on the balance of ticks and crosses, the chair of examiners specifies a 'zone of uncertainty' – illustrated here in grey. This is the zone within which the judgemental evidence suggests that the grade boundary should lie.

Table 6.1: Awarding committee judgements of script evidence

Mark	Chair of Examiners	Chief Examiner	Principal Examiner A	Principal Examiner B	Principal Examiner C	Principal Moderator
54	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓✓✓
53	✓✓✓	✓✓✓	✓✓?	✓?✓	✓?✓	✓✓✓
52	✗?✗	✓✗✗	✓✗✓	✓✗✓	✓✗✗	✓✗✗
51	✓✗✗	✓✗✗	✗✗?	✗✗✗	✗✗✗	✗✗✗
50	✗✗	✗✗	✗✗	✗✗	✗✗	✗✗

Once script scrutiny is complete, the chair of the awarding committee weighs up the statistical and judgemental evidence available and recommends the final grade boundary, taking into account the advice of the awarding committee and exam board technical experts. Grade boundaries are set in this manner on each unit for the 'key' grade boundaries (A and E), and the remaining boundaries – A*, B, C and D at A level – are calculated arithmetically. Student performance on each unit is aggregated together to give the final qualification grade, although the method for this differs depending on the structure of the qualification.

When setting grade boundaries, the chair of examiners must consider the overall qualification level outcomes, since they are compared to the statistical predictions and are subject to reporting tolerances applied by the exams regulator (Ofqual, 2017). The reporting tolerances specify the range within which the outcomes at grade A in each subject would be expected to fall relative to the statistical predictions and are based on the size of the cohort. For example, where the predictions include over 3,000 students, outcomes are not expected to deviate by more than 1 per cent from the prediction, whereas in cases involving only 500–1,000 students, the outcomes are not expected to deviate by more than 3 per cent from the prediction.

The grade boundaries recommended by the chair of examiners are then submitted to the responsible officer of the exam board, who has overall responsibility for the decisions. The responsible officer reviews the outcomes, considering any issues that the awarding committee has raised and taking account of external information such as results in other subjects and results in the same subject from other exam boards (the exchange of data is facilitated by Ofqual). The grade boundaries can be moved at this point but with the chair of examiners' agreement and not usually outside the 'zone of uncertainty'.

Before any results are issued, exam boards' outcomes in each subject are reviewed by Ofqual, using the tolerances discussed earlier. Where the results are out of tolerance, the exam board has to provide justification to Ofqual. This can be based on additional statistical and/or judgemental evidence. Each year Ofqual has accepted some out of tolerance explanations and those results have stood (e.g. see Ofqual, 2017). On other occasions, Ofqual has challenged the explanations if they have not been supported by sufficient evidence and the proposed outcomes have been changed.

A level results are reported to students in mid-August each year and national outcomes are reported extensively in the media. Results are also sent directly to UCAS to finalize university admissions.

Definition of standards

The process described above for maintaining standards for A levels (and GCSEs) in England relies on a combination of statistical and judgemental evidence – primarily, statistical predictions and examiners' qualitative judgement of students' work. This approach to maintaining standards has been described as weak criterion-referencing (Baird *et al*., 2000) or, more recently, attainment-referencing (Newton, 2011). While both statistical and judgemental evidence is used when setting grade boundaries, the balance of evidence that exam boards prioritize has shifted over recent years (Newton, 2011). Research highlighting potential biases in examiner judgement and the tendency for examiners to give students the 'benefit of the doubt' (see Baird, 2007) has led to greater emphasis on the statistical evidence (Baird and Gray, 2016). Furthermore, the introduction of the comparable outcomes approach by Ofqual has brought the statistical evidence to the fore. Thus, while the basic principle of the standards maintaining process is to retain a level of performance from one year to the next, this is largely achieved through the use of statistical predictions.

The comparable outcomes approach is rooted in research by Cresswell (2003) into setting standards in examinations when a revised syllabus is introduced. The basic premise is that if the nature of the cohort sitting a qualification each year does not change, then the outcomes should not change either. The approach therefore prioritizes comparable outcomes rather than comparable performance (though in a period of stability comparable outcomes and comparable performance should be aligned – comparable performance would prioritize student performance on the assessment rather than the outcomes that they achieved). One reason for prioritizing comparable outcomes is that it protects students taking their assessments in the first year of a new qualification, when teachers and students are less familiar with the assessment and performance is likely to dip (Ofqual, 2016a). The alternative approach, prioritizing comparable performance, would likely result in a drop in outcomes in the first year of a new qualification, then rise over time as teachers and students became more familiar with the assessment. This would introduce unfairness into the process, since the grades that students achieved would be influenced by the point at which they sat assessments within the lifetime of a qualification.

The comparable outcomes approach is statistically driven in that prior attainment-based predictions guide the awarding process, and it is against these predictions that outcomes are evaluated by Ofqual. Within

the framework for defining standards outlined by Newton (2011), the comparable outcomes approach therefore implies a causal definition of standards, since it is the causes of attainment – in this case students' prior attainment – that one would expect to be similar for those achieving similar grades. As such, students achieving similar A level grades would be expected to have similar inputs to their learning (in this case prior attainment). Despite this, there is scope within the standard setting process for other evidence – including evidence about students' performance – to be provided as a justification for outcomes that do not align with the statistical predictions. This means that, in practice, standard maintaining can rely on a combination of both statistical predictions and judgemental evidence – an approach described as attainment-referencing (Newton, 2011). Since attainment-referencing refers to both statistical and judgemental evidence, it has been argued that this approach implies both a causal and a phenomenal definition of standards (Baird and Gray, 2016). This does not fit neatly within the framework proposed by Newton (2011) and raises theoretical issues for the definition of standards (Baird and Gray, 2016).

Public controversies about A levels (and GCSEs)

Results from high-stakes public exams in England, including GCSEs and A levels, are subject to intense public and media scrutiny each August when the results are issued. Prior to the introduction of comparable outcomes, GCSE and A level outcomes typically rose year on year (Ofqual, 2015). This led to various assertions that exam standards were falling; exams were getting easier; more students doing well must be a bad thing; and that increased participation and success would lead to poorer standards (Murphy, 2004). Such claims typically played out in the media in what has been described as the 'silly season' – newsrooms struggled to find newsworthy items during the quieter summer months, leading to an inevitable focus on exam results, a topic that is of relevance to a large proportion of the population (Warmington and Murphy, 2004). Such media coverage is damaging, since it can undermine public confidence in the exam system and the grades that students achieve (Simpson and Baird, 2013).

The media and public debate around falling exam standards was frequently associated with claims that exams were being 'dumbed down'. Commentators argued that the assessments no longer reflected the standard that they once did and that this had resulted in the increase in outcomes. A number of factors were cited as contributing to this,

including changes to the structure and content of the assessments, but one key factor that gained traction publicly related to the choice of exam boards. While not unique, the assessment system in England is unusual in that there have always been multiple exam boards offering qualifications in the same subject. This means that exam boards, registered charities or profit-making organizations operate in a competitive market, and schools and colleges can choose between providers. These arrangements led to claims that exam boards were lowering their standards to boost pass rates, with the aim of increasing their market share (and therefore their income). Essentially, exam boards were accused of competing on exam standards rather than on their products, raising concerns of a 'race to the bottom'. Such claims, though rejected by the exam boards, eventually came to a head in 2012 when the then Secretary of State for Education, Michael Gove, announced plans to replace the GCSE qualification with an English Baccalaureate Certificate and have a single exam board for each subject. This was to be based on a competitive bidding process. A single exam board was heralded as a way to end the year-on-year increases in exam results (DfE, 2012), although evidence from other jurisdictions suggests that this might not necessarily be the case. For example, while there is a single exam board in Scotland, the proportion of students there achieving the top grades for the Higher exam has still increased over time (see Baird and Gray, 2016).

Ultimately, the plan for a single exam board for GCSEs never came to fruition and was withdrawn a few months later. However, the presence of multiple providers for each GCSE and A level subject does raise issues around exam standards and the importance of ensuring comparable standards, in this case between exam boards. Students use their GCSE and A level grades to compete against one another, meaning that to ensure fairness, a grade A in a particular subject from one exam board must be of the same standard as a grade A in the same subject from another exam board.

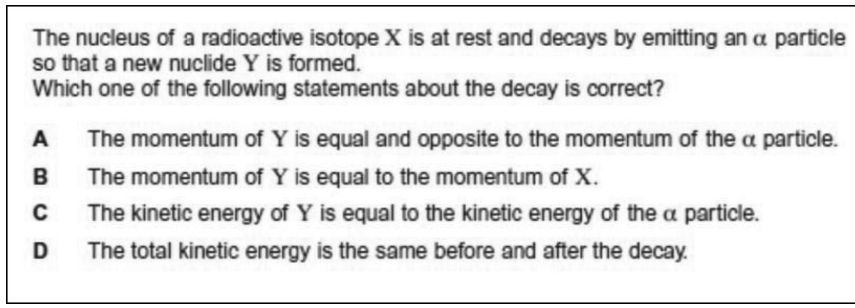
More recently, comparable standards between exam boards have been promoted through the comparable outcomes approach, since all exam boards use the same statistical evidence to guide their awards and outcomes are reviewed by Ofqual. Prior to this, exam boards used statistical evidence to support the setting of grade boundaries, but this was not necessarily done in a consistent manner. The intentions of introducing comparable outcomes were therefore two-fold: to promote comparable standards between exam boards and to protect students when changes were made to the assessments.

Since the introduction of comparable outcomes, GCSE and A level outcomes have largely remained stable (Ofqual, 2015). This has proved fairly uncontroversial at A level, yet there has been greater criticism of the approach at GCSE. This reflects the different uses to which exam results are put. GCSE results are key to accountability measures against which the performance of schools is judged, while A levels are primarily used for selecting students for higher education. As such, provided that A levels can differentiate between students effectively, outcomes being stable over time is likely to be less of a concern.

The main criticism now levelled at comparable outcomes for GCSE is that by using statistical predictions based on prior attainment, the approach effectively caps outcomes and does not allow genuine improvements in student performance to be recognized. Consequently, although schools are under intense pressure to improve results and resources are channelled towards this, outcomes are not able to increase in response. Since introducing comparable outcomes, Ofqual has always stated that there is scope for exam boards to provide evidence to support outcomes that do not align with the statistical evidence, though in practice such cases are relatively rare (Ofqual, 2017). Furthermore, generating compelling evidence that demonstrates genuine improvements in performance, rather than increasing familiarity with the assessment, is far from straightforward, an issue acknowledged by the regulator (Ofqual, 2015). In recognition of this, Ofqual has committed to researching methods of improving awarding in its corporate plan (Ofqual, 2016b) and has introduced a national reference test from 2017. The national reference test is intended to monitor any changes in performance in English and mathematics by 16-year-olds and may be used in GCSE awarding from 2019, though the technicalities of how this may operate are still under discussion.

The case of comparable outcomes provides some insight into the way in which exam standards are perceived publicly in England. Prior to the introduction of the comparable outcomes approach, rises in exam results year on year were frequently cited as evidence that exams standards were falling and that exams were getting easier. Now there are claims that a comparable outcomes approach is preventing outcomes from rising to recognize improvements in student performance at GCSE. Given the methods used to maintain standards over time in these high stakes qualifications, it is unlikely that they will ever be immune to challenge from the public, the teaching profession and politicians. That puts an onus on researchers and regulators to progress thinking about how standards are defined and to improve the processes used to maintain standards.

Annex 1. Examples of A level exam questions



The nucleus of a radioactive isotope X is at rest and decays by emitting an α particle so that a new nuclide Y is formed.
Which one of the following statements about the decay is correct?

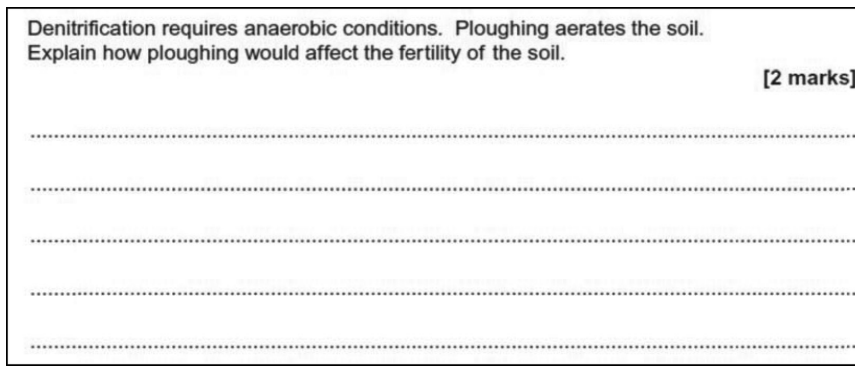
A The momentum of Y is equal and opposite to the momentum of the α particle.

B The momentum of Y is equal to the momentum of X.

C The kinetic energy of Y is equal to the kinetic energy of the α particle.

D The total kinetic energy is the same before and after the decay.

Figure 6.1: AQA A level Physics multi-choice question (2015)



Denitrification requires anaerobic conditions. Ploughing aerates the soil.
Explain how ploughing would affect the fertility of the soil.

[2 marks]

.....

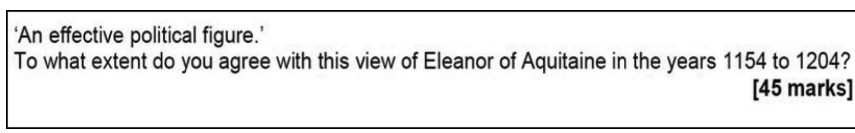
.....

.....

.....

.....

Figure 6.2: AQA A level Biology short-response question (2015)



'An effective political figure.'
To what extent do you agree with this view of Eleanor of Aquitaine in the years 1154 to 1204?

[45 marks]

Figure 6.3: AQA A level History long-response question (2015)

References

Baird, J., Cresswell, M. and Newton, P. (2000) 'Would the real gold standard please step forward?'. *Research Papers in Education*, 15 (2), 213–29.

Baird, J. and Gray, L. (2016) 'The meaning of curriculum-related examination standards in Scotland and England: A home–international comparison'. *Oxford Review of Education*, 42 (3), 266–84.

Baird, J. (2007) 'Alternative conceptions of comparability'. In Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority, 124–65. Online. https://goo.gl/8SvTBo (accessed 7 June 2018).

Cresswell, M.J. (2003) *Heaps, Prototypes and Ethics: The consequences of using judgements of student performance to set examination standards in a time of change*. London: Institute of Education.

DfE (Department for Education) (2012) 'Consultation on KS4 reform'. Online. http://webarchive.nationalarchives.gov.uk/20121102154825/http://media.education.gov.uk/assets/files/pdf/r/reforming%20key%20stage%204%20qualifications%20-%20consultation%20document.pdf (accessed 27 July 2018).

DfE (Department for Education) (2016) 'Provisional A level and other 16–18 results in England, 2015/2016'. Online. https://goo.gl/uXRZzo (accessed 9 June 2018).

DfE (Department for Education) (2017) *Revised Destinations of Key Stage 4 and Key Stage 5 Students, England, 2014/15*. London: DfE. Online. https://goo.gl/DstGxM (accessed 9 June 2018).

DfES (Department for Education and Skills) (2005) *14–19 Education and Skills White paper*. Online. https://goo.gl/k6YuwG (accessed 9 June 2018).

JCQ (Joint Council for Qualifications) (2017) *A, AS and AEA results, Summer 2017*. Online. www.jcq.org.uk/examination-results/a-levels/2017/main-results-tables (accessed 9 June 2018).

Murphy, R. (2004) *Grades of Uncertainty: Reviewing the use and misuses of examination results*. A report commissioned by the Association of Teachers and Lecturers.

Newton, P.E. (2011) 'A level pass rates and the enduring myth of norm-referencing'. *Research Matters*, Special Issue 2: 20–6.

Ofqual (Office of Qualifications and Examinations Regulation) (2015) *Setting GCSE, AS and A level Standards in Summer 2014 and 2015*. Online. https://goo.gl/oTNpJ8 (accessed 9 June 2018).

Ofqual (Office of Qualifications and Examinations Regulation) (2016a) *An investigation into the 'Sawtooth Effect' in GCSE and AS / A level assessments*. Online. https://goo.gl/DPZgwZ (accessed 9 June 2018).

Ofqual (Office of Qualifications and Examinations Regulation) (2016b) *Corporate Plan 2016–2019*. Online. https://goo.gl/LkXFLX (accessed 9 June 2018).

Ofqual (Office of Qualifications and Examinations Regulation) (2017) *Summer 2017 GCSE, AS and A Level Awards: A summary of our monitoring*. Online. https://goo.gl/nCd5q5 (accessed 9 June 2018).

ONS (Office for National Statistics) (2016) *Population Estimates for UK, England and Wales, Scotland and Northern Ireland: Mid-2016*. Online. https://goo.gl/ces4By (accessed 9 June 2018).

Simpson, L. and Baird, J. (2013) 'Perceptions of trust in public examinations'. *Oxford Review of Education*, 39 (1), 17–35.

UCAS (Universities and Colleges Admissions Service) (2017) *End of Cycle Report 2017*. Online https://goo.gl/G14j3r (accessed 9 June 2018).

Warmington, P. and Murphy, R. (2004) 'Could do better? Media depictions of UK educational assessment results'. *Journal of Education Policy*, 19 (3), 285–99.

Explaining educational standards: The challenge of uncertainty

Mary Richardson

The notion of a standard in the English education system is one that is suffused in complexity and scepticism. Since its introduction into educational settings in the 1880s (Williams, 1961), the term has become synonymous with a simplistic model that often reifies a particular practice. Such binary perceptions of how we discuss and recognize standards are unhelpful because they fail to provide a suitably nuanced discussion of both the strengths and limitations of just how standards are determined in contemporary educational contexts. A well-defined explanation and sensitive discussion of just how standard setting is conducted has the potential to challenge many of the urban myths that surround the subject. Taylor and Opposs's chapter is to be welcomed in this regard.

A striking aspect of this chapter is its depiction of the constantly changing landscape of educational policy in the English state-maintained education system. Starting with a condensed description of the present education system and then focusing on that elusive 'gold standard' qualification, the Advanced (A) GCE Level provides a conduit to the detailed descriptions of processes related to high stakes assessments. A clearly written guide to the processes involved in awarding is long overdue. The examination boards in England provide guides, but stakeholders (e.g. teachers, parents or examination officers) might believe that boards differ in their approaches to standard setting, and this negates the reality that practice is carefully regulated. Educational standards debates have intensified in England since the late 1980s and the introduction of a national curriculum and nationally reported testing in schools. Stricter codes of accountability in schools have also added fuel to concerns about standards, and public discourses in 2018 are characterized by fast-paced comment in public spaces via the internet and, more specifically, in the more personal realms of social media.

Debate about education is to be welcomed, but it is important to be cognisant of the standard of the arguments that frame such debates. Too often, as Murphy (2013) argued, memorable headlines are not always built on strong foundations; opinion, anecdotal experience and personal belief often

underpin the claims. It is the detail that matters in the processes of standard setting in education, yet the facts are often overlooked, misunderstood or simply ignored when standards are discussed in public domains such as print media, or online via social/news media. Taylor and Opposs acknowledge the fact that the methods used by examination boards and regulatory bodies such as Ofqual cannot be 'immune to challenge', but it is crucial to state that such challenges require evidence based on fact rather than belief. Here they present an argument that is systematic in its critical examination of different awarding processes and also conscious of the inherent lack of 'one perfect way' to determine a standard. Opportunities to have open, public discussions about the complex nature of standards in England are vital in sustaining trust in awarding systems. Such difficult conversations are worthwhile because the trust they inculcate then underpins the value of our education system and endorses the importance of being educated.

References

Murphy, R. (2013) 'Media roles in influencing the public understanding of educational assessment issues'. *Oxford Review of Education*, 39 (1), 139–50.

Williams, R. (1961) *The Long Revolution*. London: Chatto and Windus.

Ensuring standards in English A levels

Peter Tymms

Taylor and Opposs provide a clear, and necessarily condensed, account of what is a unique standard setting system in England, with a focus on A levels. This has, as they point out, evolved over many years, and a more detailed historical account can be found in Tattersall (2007). In contrast, the international systems, such as PISA, TIMSS and PIRLS, are relatively recent. These systems broadly differ from A levels in that they have tended to rely on item response theory in their analyses, employ pre-testing and use objective item formats.

Restriction of space has meant that Taylor and Opposs were only able to provide an overview. This short commentary will pick up just three points from their chapter that may seem puzzling to an international audience, and expand on them.

Some readers might wonder why England goes to such trouble to produce A level exams when much quicker and cheaper alternatives such as the Scholastic Assessment Test (SAT) (Kobrin *et al.*, 2008) are available and have been widely used for college entrance in the US. There are three answers to this. The first is tradition, which can be hard to alter. The second is that the existence of A levels provide meaningful motivation for students and their teachers in schools and colleges to work hard at their chosen subjects. The third is that A level results are better predictors of university success than the SATs (Kirkup *et al.*, 2010).

Grade inflation at A level is one of the major reasons why there has been such a serious focus on standards in recent years. Although some newspapers have made hay with the evidence (if it bleeds it leads), grades at all levels of national tests and exams in England have exhibited grade inflation over a long period and this resulted in the top grade at A level, an 'A', being given to such a high proportion that a new 'A*' grade had to be introduced. During the 1990s, the high proportion being awarded the top grades made it impossible for the most selective universities to discriminate among the more able students and this was one of the reasons why action was needed. At the same time, exam boards and the qualifications regulator were becoming aware of issues around inter-board comparability. The comparable outcomes approach was developed to deal with this issue. We

can be pretty sure that there was grade inflation, even though there are competing explanations, because the Centre for Evaluation and Monitoring (CEM) had been accumulating extensive data over many years using the same test year on year and these tests could be used as anchors (Tymms *et al.*, 2005; Tymms, 2004; Coe, 2007; Coe and Tymms, 2008). When the comparable outcomes approach became a regulatory requirement, it addressed inter-board comparability and had the additional beneficial effect of ending grade inflation.

The third point concerns the existence of several exam boards. An initial reaction to issues associated with standards is to argue for a single awarding body. But there are advantages to diversity. They include competition, which can encourage innovation. Such innovation can be positive, in which case, all boards can eventually adopt the new ideas. It may, alternatively, have a negative impact in which the new ideas are restricted to a single innovative board. Diversity is also valuable in ensuring that choice of syllabus is available. This can allow well-qualified teachers and lecturers to select an alternative to match, as closely as possible, their preferred content when teaching pre-university courses.

In summary, the chapter entitled 'Standard setting in England: A levels' provides an excellent overview of the A levels standard setting system in England. This commentary discusses just three points that may appear to be odd to anyone not familiar with the system. The first concerns the use of curriculum focused tests, which include extended answers, rather than cheaper, quicker multiple choice tests of ability. Briefly, A levels are better predictors and encourage better pedagogy. The second relates to grade inflation, which is often an issue in grading systems. It came to a head when universities were no longer able to select the most able students as so many were being awarded the top grades. This probably provided the impetus for the comparable outcomes approach discussed in the chapter. Finally, some arguments are presented to defend the existence of more than one awarding body. They emphasize the value of diversity for competition, innovation and education.

References

Coe, R. (2007) *Changes in Standards at GCSE and A-Level: Evidence from ALIS and YELLIS*. Report for the ONS, CEM (Curriculum, Evaluation and Management) Centre, Durham University. Online. https://goo.gl/oKgwkB (accessed 10 June 2018).

Coe, R. and Tymms, P. (2008) 'Summary of research on changes in educational standards in the UK'. In Harris, M. *Education Briefing Book 2008*. IoD Policy Paper. London: Institute of Directors, 86–109.

Kirkup, C., Wheater, R., Morrison, J., Durbin, B. and Pomati, M. (2010) *Use of an Aptitude Test in University Entrance: A validity study*. London: Department for Business, Innovation and Skills. Online. https://goo.gl/3wvFvt (accessed 10 June 2018).

Kobrin, J.L., Patterson, B.F., Shaw, E.J., Mattern, K.D. and Barbuti, S.M. (2008) *Validity of the SAT® for Predicting First-Year College Grade Point Average*. Research Report No. 2008–5. New York: College Board. Online. https://goo.gl/24sFzm (accessed 10 June 2018).

Tattersall, K. (2007) 'A brief history of policies, practices and issues relating to comparability'. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards*. QCA, 43–96. Online. https://goo.gl/yiVDit (accessed 10 June 2018).

Tymms, P. (2004) 'Are standards rising in English primary schools?'. *British Educational Research Journal*, 30 (4), 477–94.

Tymms, P., Coe, R. and Merrell, C. (2005) 'Standards in English schools: Changes since 1997 and the impact of government policies and initiatives'. A report for the *Sunday Times*. Durham: University of Durham. Online. https://goo.gl/GNTw96 (accessed 10 June 2018).

Chapter 7

Standard setting in France: The baccalauréat

Roger-François Gauthier

Introduction

France: A centralized system of certification that does not prevent paradoxical organization

The baccalauréat is a very sensitive topic in France, for the government and the public. The main political provisions concerning it have mainly been quantitative during the three last decades, that is, they have focused on the proportion of young people expected to pass the baccalauréat: the Orientation Law of 1989 stipulated that 80 per cent of the relevant age group (all young people who are 18 years old each year) should reach baccalauréat standard (*le niveau du baccalauréat*) before 2000. Indeed, every year, about 80 per cent of each age group (695,682 students in 2016) sit this examination. The success rate of those sitting the exam was 88.6 per cent in 2016; in total, more than 65 per cent of the age group pass it successfully. About 80 per cent of students who have successfully passed the examination (i.e. mainly students from the academic and technological streams) immediately enter higher education. During the last five decades, the baccalauréat has evolved quantitatively and structurally in that two new types of baccalauréat were added with the creation of technological and vocational baccalauréats.

The baccalauréat has the double function of a school leaving examination and of a first university grade. Although there are three main streams (*voies*) of baccalauréats (academic, technological and vocational) and many specializations within each of these three streams (about one hundred for the vocational baccalauréat for instance), the baccalauréat by itself gives the right to enter any university regardless of the specialization of the candidate. The Ministry of National Education is directly responsible for school leaving examinations, and the bodies running the examinations are under its direct responsibility. The position of the state is monopolistic, as it has to design the various streams of schooling, to stipulate the curricula

as well as the ways students are assessed, to lead the assessment process every year and to evaluate the whole organization of senior schooling.

A specific issue in France springs from the fact that higher education is divided into two main parts: universities that just request the baccalauréat, and the *grandes écoles*, both public and private, that are highly selective. To be able to sit for the competitive examinations that open access to these *grandes écoles*, students have first to be admitted for two years to so-called 'preparatory classes' (or vocational courses leading to a bac+2 diploma, i.e. a diploma obtained after studying for two years post-baccalauréat). They have to apply for admission to these preparatory classes during the final year of high school, which is the year of the baccalauréat. What is at stake here is much more important than passing the baccalauréat (where success rates are so high that there is no question about the fact that they will pass it successfully). Admission to these preparatory classes is not based on the results of the baccalauréat, namely a national and anonymous examination, but only on school-based assessment, which is almost entirely excluded from the baccalauréat (see below).

Baccalauréat, both a national school leaving examination and a university entrance examination

Although the term 'baccalauréat' had been used previously, the modern baccalauréat was created in 1808, to be both the first degree in higher education and the leaving certificate from secondary education. Since its creation, when there was only one 'stream', there has been a process of double, and opposing, developments:

1. Progressive diversification into what are currently called the three *voies* (routes, i.e. academic technological and vocational), *séries* (secondary streams, three for the *academicoute*, but much more for the others) and 'specialities'. For instance, scientific baccalauréat students must, in addition to courses followed by all students of the scientific *série*, choose either a maths, physics-and-chemistry, biology-and-geology, technological sciences or computer sciences specialization.
2. Progressive merging of many *séries*, for two reasons:
   - technological and vocational training requires less specialization and more transferable competences than before;
   - governments have tried to block unwanted and negative effects of the appearance of a social hierarchy between some *séries* by merging them.

Generally, students sit the baccalauréat at the end of the two last years of *lycée* (senior high school that lasts three years). Most students are 18 when they sit the baccalauréat. Just one examination is taken, covering the various subjects of each *voie* and *série*: each baccalauréat requires around eight to twelve papers and oral examinations, which differ from one *série* to the other and between which there is no choice, except for additional and optional subjects. Some papers and oral tests are taken one year before the end of *lycée* (French literature, for instance, is commonly not taught during the last year, leaving room for the compulsory teaching of philosophy), but most of them are taken at the end.

Students choose which baccalauréat they are going to sit in two steps: when they enter high school (three years before the exam) they choose their *voie*, and after one year they choose their *série*. It is not a free choice for the student, as the final decision is up to the junior or senior high school principal. The students express their wishes, but the teachers and the principal always have the final decision.

Students pass the examination and get the baccalauréat if they get an overall average mark equal to or higher than 10 out of 20 after addition of all the results obtained in all the subjects (papers and oral tests) prescribed in each *série*. Each subject is affected by a weighting factor varying from one *série* to the other. This permanent weighting factor has been fixed by law for each *voie* and *série* to which it contributes. For example, a student in the scientific stream has a compulsory test in history, with a weighting factor much weaker than a student in the *littéraire* stream, where French, philosophy, languages and history have the highest weighting factors. Students whose overall average mark is under 8/20 fail; those marked between 8/20 and 10/20 sit two resit tests or oral retakes. The student chooses the topics in which he/she wants to resit in order to better his/her mark and is reassessed a few days after the initial results. The overall average result is recalculated after substitution of the two new marks.

Obtaining the baccalauréat, whether academic, technological or vocational, entitles any student to enter any university in any subject. Although necessary, the baccalauréat does not entitle students to enter elite training courses (*classes préparatoires*) leading to *grandes écoles*, or selective vocational courses leading after two years to a bac+2 vocational diploma.

Changes to the baccalauréat have been few and not fundamental in recent years. The most recent reform mainly relates to the vocational baccalauréat, created in 1985. Since 2010 it requires the same number of years of schooling in a vocational senior high school (three, rather than the

four years required between 1985 and 2010) as do general and technological baccalauréats.

Another reform dates from 2016: the candidates who fail are allowed to save their individual subject paper marks for the next five exam sessions, if these marks are equal to or above 10 out of 20. They have to resit only the subjects in which they scored below this mark. We still do not know the effect of this decision.

There have been few reforms of the baccalauréat over the years since the issue is controversial among policymakers, who generally prefer not to run the risk of introducing change.

The assessment process

The assessment process described below has remained unchanged in its main characteristics for decades, except for the vocational baccalauréat that was created in 2005, where a new form of assessment called '*contrôle en cours de formation*' (standardized tests organized several times during the year instead of a single final test) was introduced.

About 75 per cent of the marks are given for written anonymous exams, the weight of oral examinations being about 10 per cent of the marks, the weight of school-based assessment being limited to 5 to 10 per cent (physical education mainly) and the weight of a personal work to be orally presented being limited to 5 to 10 per cent. Most written tests consist of *dissertations* (in history, philosophy, French literature and social sciences). Multiple choice assessment is absent.

A typical written test lasts four hours. A candidate will have about twenty-four hours of assessment for the whole examination in a short period of time. There are no resits except in the restrictive conditions previously explained.

There is just one marking system, marks from 0/20 up to 20/20. There are no 'pass marks' for the various subjects, as all of them are included in an average overall mark, with 10/20 as a pass mark. As already noted, this average mark is calculated with weighting factors that for the same subject can vary from one *série* to another.

There are several steps in the way test requirements are elaborated:

- The way a subject is assessed is defined in a permanent rule (*définition d'épreuve*) established by the Minister of Education. In this *définition d'épreuve* one can find what each assessment should look like. The *définition d'épreuve* is elaborated by the *Inspection générale de*

l'éducation nationale (a body divided into groups dedicated to the various subjects, and, placed under the authority of the minister) and published as an official rule by the Ministry of Education;
- The content is permanently defined in the curriculum itself (*programme d'enseignement*), although the link between it and the *définition d'épreuve* is not always clear;
- Annually, the various tests are developed in the name of the minister. Often one assessment test is available for the whole national territory. Two different persons are appointed to develop the test: one is an *inspecteur général* of the subject to be assessed, belonging to the ministry, the other a university teacher; they often chair commissions composed of practising teachers, who make proposals. The involvement of the university teacher is often more formal than real. Before being chosen, a test is trialled by 'guinea pig' teachers who have about half the time of the students and who have to write a report about the feasibility of the test. The final cut is a ministerial decision. As to any quality assurance process for production of assessments, there does not appear to be anything of this kind, although some regions check the organization of the test (how it is protected from potential leaks, for instance).

New assessments have been implemented in recent years, intending to enrich the assessment of some subjects (e.g. oral assessment in foreign languages) or to check that all subjects are assessed (e.g. sciences in non-scientific streams):

- For foreign languages, and in conformity with the European framework, two foreign languages are now assessed, not at the end of the year but during the school year, through written as well as oral tests. Previously only written tests existed (shorter and less expensive to organize).
- In 2011, a test in sciences covering a mix of biology, physics and chemistry was introduced one year before the final year for non-scientific students.

As to the marking system itself, marking (i.e. giving a mark on a scale from 0 to 20, with the possibility of quarters and halves of marks) is done by selected teachers (who will not have taught the students whose papers they have to mark). These teachers receive the papers at home and have to mark them within a limited number of days. They usually enter their marks on a computerized system.

Depending on the subject, the time needed for marking is one of the factors that determines the duration of the examination in the great number of assessed subjects: marking takes at least three weeks at the end of each year. As the high schools are often totally requisitioned for the organization of the examinations, one could say that because of the examination students are deprived of about ten weeks of teaching out of the three years of *lycée* schooling.

The markers receive two kinds of help to mark the papers:

- One is a *barême*, or marking scale, that says what part of the mark has to be attributed to the various parts of the test assessment. This marking scale is more effective in some subjects (e.g. maths) than in others (e.g. philosophy). For the subjects where a marking scale is provided, the *barême* gives the expectations for a 5 or a 10 in that specific test. For the subjects where no marking scale is provided, the expectations are implicit and supposed to be part of the professional know-how of the teachers.
- The other is the existence of *commissions d'harmonisation* (one for each subject), namely groups of experienced teachers and inspectors from the local level (these inspectors are appointed at a regional level, and differ from *inspecteurs généraux*, who report only to the minister), that will join markers during the marking process in order to help adjust and to some extent standardize the marking.

Standard setting process

There is no standard setting process; there is just marking. The exam leads to pass/fail decisions: the various marks attributed by the markers in all the subjects are collected by the local jury, composed of all markers from all subjects involved in the marking. Each jury is dedicated to a number of candidates and is sovereign for those candidates' results: it makes the final decisions. Each jury is chaired by a university teacher (whose specialization does not matter) and gathers the results of all the subjects. As in the test development process, the chairing of the juries by university teachers is more formal than real. The inspectors present in the juries often seem to play a deeper role.

Each jury meets twice: once after the marking of all subjects, and then again after the resit tests for candidates whose average mark after weighting is between 8 and 10 out of 20. The jury takes two decisions: whether a candidate passes or fails and, if he or she passes, whether a *mention* (average,

satisfactory, good, excellent) is attributed to the candidate, still on the basis of the average mark.

The decision is in fact made by the computer (from the average mark) for most students; the jury can discuss only the borderline cases and can in these cases use the *livret de baccalauréat*, including the results obtained by the student during the two previous years. This use of school-based assessment is in fact very limited, despite the fact that this *livret* gives the level of each student in four major competences for each subject.

Appeals are possible but can only check the absence of any material error: when the judges consider appeals, they shelter behind the sovereignty of the jury and refuse to reconsider any mark. For this reason, the number of appeals considered by judges is low.

Political and public controversies about the baccalauréat

It can be argued that there are currently no real public controversies about the baccalauréat or the examination standards. Each year in July the successive ministers are glad to publish the success rate of the examination as it is both high and increasing almost every year (a steady increase from 64 per cent in 1984 to 88.5 per cent in 2016). They comment on it as evidence of the quality of teaching and learning in high schools.

One might think that the high failure rate of students during the first two years of higher education should raise questions about baccalauréat standards, but surprisingly that does not occur: the usual political trend is more to question higher education itself ('Why do so many students fail?') or the fact that when they have got any kind of baccalauréat, students can freely enrol in any university for any specialization.

Since 1984, the main political issue about the baccalauréat has been the quantitative objective of ensuring that by 2000 80 per cent of the age group should sit the baccalauréat. This objective was adopted by law in 1989, and broadly speaking has been achieved, although with a long delay (the percentage of the age group sitting the examination in 2016 was 78.6 per cent). It remains a purely quantitative objective, the issue of standards having never been called into question. The idea still prevails that the prescribed curriculum, together with the various traditions of the various topics and subjects, are enough to preserve good standards. Several hypotheses can be made about this weak interest in the clarification of standards:

- Maybe the old tradition of the subjects in French secondary education has long been content with a conception of knowledge that favours

each student's freedom more than requesting competences and skills easily standardized; the idea paradoxically is that standardizing could weaken a level of achievement that, by the way, is not known or measured.

- Maybe the political powers value the social meaning of ensuring the bulk of the population passes the baccalauréat and does not want to better know its epistemological meaning. Since the baccalauréat is not high stakes for the French elites, one can ask whether it is worth taking the trouble and opening Pandora's Box.

The standards are not known, the examination is not independently evaluated and nobody seems to care. An interesting point to illustrate the fact that the baccalauréat is a blind spot in the French educational system is the scarcity of research work about both the baccalauréat and the standards: the question of standards has never systematically come up, neither from a sociological nor from an epistemological point of view. When asked why so few research works are dedicated to the baccalauréat, some researchers answer that because of a strong social resistance there is no chance of introducing any change in the current examination.

In regard to the political and public views of examination standards, it can be assumed that, up to the present time, both policymakers and the public accept a relative ignorance about what the baccalauréat as a whole checks and proves. Its formal and juridical meaning – that is, the first grade in higher education, even if it is largely a fiction – still seems an adequate reference.

We could interpret this situation as the long-term consequence of a centralized educational system that believes in itself and in its traditional strength. Obviously this self-confidence has not existed for at least two decades for compulsory education, partly as a consequence of PISA tests. For this level (compulsory education), since 2005 the French educational system has invented a totally new paradigm with the '*socle commun de connaissances, de compétences et de culture*'. But for *lycée* (secondary high school) level, and the baccalauréat, nothing similar has been put in place.

The question of the *lycée* and its possible evolution tends to be avoided by policymakers. The reasons for this abstention have not been studied enough, but we have mentioned that the transition between secondary and higher education exists in tension between an official position and the reality. In the official and traditional position, this transition is transparent and fair, through an anonymous examination. The reality is more hypocritical, as the access to the best channels of higher education (preparatory classes for

grandes écoles) follows other roads than a fair examination, through the use of school-based assessment, without any quality control on its validity.

This hypocritical system is well understood by families who have the keys and rules of this social game: it is not in their interests to change anything in the baccalauréat organization. Nevertheless, some policymakers have recently introduced a proposal to consider schooling at *lycée* in the bigger framework of what is called the 'bac-3/+3' issue: according to them it is necessary to think of the *lycée* as a part of a larger system, beginning three years before the baccalauréat and ending with a licence at a bac+3 level.

In January 2018, after the completion of this case study, the new minister of education Jean-Michel Blanquer announced a reform of *lycées* and the baccalauréat. To a certain extent one could say that these reforms, which are based on a report by Pierre Mathiot (2018), address some of the criticism presented above. These political decisions, which will result in a new baccalauréat in 2021, aim to return to the baccalauréat a clear meaning as well as a real function in the educational journey of students. They consist, at least for the *baccalauréat général* (the *baccalauréat professionnel* is not concerned and the *baccalauréat technologique* only partially) of the following changes:

- removing the *séries* described above, meaning that students have to choose between more in-depth specialized courses
- continuing the obligatory teaching of academic culture for all, which will include the traditional disciplines of *lycées*, including philosophy
- simplifying the final examinations by limiting them to five subjects, four of which will be taken in the final year of school: French, the two courses (*spécialités)* chosen by the student, philosophy and an interdisciplinary *grand oral*
- introducing continuous assessment in the form of national tests
- ensuring that the content prepares students better for the requirements of higher education. The results of examinations will be partially included in applications for higher education.

The baccalauréat, which continues to be designed as a collection of disciplines, will thus develop a clearer function. One question will be whether families and students, particularly the elites, who value a versatile baccalauréat (the current scientific *série*) to keep their options open until the entry to higher education will accept specialization at the end of the first year of *lycée*.

The baccalauréat: From elite selection to mass certification

Jean-Pierre Jeantheau

How can it be that Chloé, a Réunion Island candidate in the 2017 examinations for France's iconic school-leaving diploma, the baccalauréat (well described by Roger-François Gauthier), could achieve a final mark of 21.29 out of 20 (Bariéty, 2017)? In effect, the final mark is a weighted average of all achieved subject examination marks, each subject weighted by a pre-determined coefficient reflecting its importance in the specialist diploma concerned (e.g. mathematics is weighted 7 and history-geography 3 in the *baccalauréat scientifique*). But additional elective tests are available, and where marks higher than 10 are achieved in these they are weighted appropriately and added to those obtained in the compulsory tests. This leads to an increase in the size of the numerator used to produce the weighted average mark. The weights associated with the elective subjects, however, are not added into the denominator, which therefore remains unchanged. The overall result is a higher weighted average mark: for example, for the *baccalauréat section Sciences/Sciences de l'Ingénieur/spécialité Informatique et Sciences du Numérique*, the maximum point (20/20) in all the compulsory subjects x coefficients = 760 (sum of coefficients x subjects = 38). Additional points with optional subjects (for instance, Greek and Horseriding) 40 points, additional points with personal work 20. Average (and maximal) mark in baccalauréat S = 760 + 60/38 = 21.58 (for a simulation, see http://etudiant.lefigaro.fr/bac/simulateur/serie-s/). This practice, like many others, is intended to help students who are weak in the compulsory subjects but have strengths in sport, foreign languages or arts. But students such as Chloé who are strong in the compulsory subjects can also benefit from the elective test availability. Hence a final mark higher than the maximum of 20 is possible.

How is this practice socially acceptable? It is because the population is divided on the meaning of the word 'success'. The government uses 'success' to designate the attainment of a pre-determined level as indicated by a diploma. The elite understand success in the competitive sport sense: getting ahead of others. The first type of success can be achieved by the masses, the second by definition cannot. Fifty years ago, passing the baccalauréat meant entering the group of 'the best', thus reconciling the two meanings of

'success'. These days, by contrast, those failing to achieve the baccalauréat are stigmatized. There is in consequence a strong social pressure to gain the diploma. For the government a significant drop in the percentages of candidates succeeding in the baccalauréat would be seen as evidence of a failure to achieve intended political objectives: in particular that by 2000 80 per cent of the age group should work towards the baccalauréat, an objective that has essentially been achieved.

These pressures are additional to budgetary concerns. The baccalauréat was estimated to cost between 50 million euros (external costs) and 1.5 billion euros (including internal costs) in 2013 (Battaglia, 2013). In this context, the baccalauréat could follow the same path as the *brevet*, the lower high school diploma, which is based in part on continuous assessment (Mathiot, 2018). Failure to obtain the *brevet* does not impede transition upwards through the school, but neither does achieving it open doors to any particular further education and training opportunities. At the same time, the need for a diploma of some kind as evidence of a degree of successful schooling motivates the weakest students to attempt to gain the *Certificat de Formation Générale* (CFG), whose demands are lower than those of the *brevet* and for which annual results do not even appear in the Ministry's statistical yearbook (Ministère de l'Éducation nationale, 2017).

French society speaks of the fight against inequality, focusing increasingly on the end rather than the means. There will always be students who are 'better' than others and, in a hierarchical society, organizations that give priority to seeking them out. The diploma, and most specifically the baccalauréat, used to be the evidence allowing the best students to be identified. It has since become an indicator used by successive ministers as evidence of improvement in the achievement of the population, even if such improvement is regularly belied by the results of international surveys, in particular PISA. And so selection of the best students is postponed by extending length of study, or achieved through the use of selection tests such as those used by the *grandes écoles*, or even private higher education institutions such as the 42 schools that do not award any end-of-studies diploma. Is playing on the meaning of the word 'success' going to be enough to meet the range of possible social demands placed on the Bac?

References

Bariéty, A. (2017) 'Chloé Tossem, meilleure bachelière de France avec 21,289 de moyenne', *Le Figaro Etudiant*, 31 July. Online. https://goo.gl/j85kNF (accessed 10 June 2018).

Battaglia, M. (2013) 'Le coût caché du bac: 1,5 milliard d'euros', *Le Monde*, 11 June. Online. www.lemonde.fr/societe/article/2013/06/10/le-cout-cache-du-bac-1-5-milliard-d-euros_3427037_3224.html (accessed 17 July 2018).

Mathiot, P. (2018) *Un nouveau baccalauréat pour construire le lycée des possibles*. Paris: Ministère de l'éducation nationale. Online. https://goo.gl/mQ31Ph (accessed 10 June 2018).

Ministère de l'Éducation nationale (2017) *Repères & références statistiques: Enseignements, formation, recherche*. Paris: Ministère de l'éducation nationale. Online. https://goo.gl/6mUhnA (accessed 10 June 2018).

Grade comparability and the French baccalauréat

Sandra Johnson

In his chapter on standard setting in France in this volume, Roger-François Gauthier offers a comprehensive, critical and highly informative overview of the baccalauréat, identifying some of the social, political and technical issues associated with this internationally recognized school leaving diploma.

From its small-scale beginning as an elitist university entrance qualification in Napoleonic times, the baccalauréat enterprise has continually grown in scale and cost, quite dramatically so during the second half of the twentieth century, as a result of politically driven reforms aimed, with limited success (Ichou and Vallet, 2011), at reducing social inequality by widening access to this nationally respected qualification (El Atia, 2008). The remodelled umbrella qualification now embraces technological (late 1960s reform) and vocational (mid-1980s reform) specialisms, alongside the academic strands of the historic *baccalauréat général* (scientific, literary, economic and social). It is available to candidates throughout metropolitan France, and its overseas departments and territories, with examination calendars and examination papers necessarily differing from one time zone to another (for example, between France and the Caribbean). To give an indication of the current scale of provision, the number of candidates who presented for examinations in 2017 in metropolitan France and its overseas departments was just under 730,000 (Thomas, 2017). Just over half the candidates presented for one or other of the three specialisms within the *baccalauréat général*, just under a fifth for one or other of the eight variants of the *baccalauréat technologique* and around a third for one or other of the many vocational variants of the *baccalauréat professionnel*. Overall pass rates were high in every case, particularly for the general and technological baccalauréats, at just over 90 per cent, and the vocational baccalauréat coming in at just over 80 per cent.

The marking of written tests is locally based (i.e. within *académies*), with marker standardization practices resembling those applied in many other countries, including the UK. Weighted average marks across all subject components determine passes and merit grades (mentions), by reference to a 0–20 legacy mark scale. Grade boundaries are fixed at 2-mark intervals, with 10 a passing mark. There is apparently no formal attempt, regionally

or nationally, through statistical manipulation or otherwise, to modify mark distributions or boundary marks in order to address any observed potential drift in attainment standards over time, or at least such practices are rarely publicly recorded (Studer and Minot, 2018: 13, offer a rare example). If pass rates and the proportions of candidates gaining mentions are genuine indicators of achievement standards, then the evidence is that standards have indeed been rising over time: an overall 75 per cent pass rate across all types of diploma in 1995 increasing to almost 90 per cent 20 years on (Thomas, 2017). But are the increases in every type of baccalauréat indicative of rising achievement? Or do they have to do with the gradual, unintended development of less difficult examination papers over time, or of relaxing marking standards? If we can say nothing about marking and grade comparability over time, or across diploma types and specialisms, what is known about potential, long-standing differences across *académies* within metropolitan France, and between these and overseas locations? Are mentions of *très bien* in the *baccalauréat littéraire* for candidates assessed in Montpellier, Lyon, Corsica or Guadeloupe equivalent, as assumed worldwide?

Both Erasmus and PISA have served to focus domestic and international attention onto the baccalauréat, with questions about utility and technical quality newly emerging. There must inevitably be injustices to students in the system – but given the continuing dearth of relevant research their nature and scale remain unknown, and fairness in assessment and future work and education opportunities left in question.

References

El Atia, S. (2008) 'From Napoleon to Sarkozy: Two hundred years of the Baccalauréat exam'. *Language Assessment Quarterly*, 5 (2), 142–53.

Ichou, M. and Vallet, L.-A. (2011) 'Do all roads lead to inequality? Trends in French upper secondary school analysed with four longitudinal surveys'. *Oxford Review of Education*, 37 (2), 167–94.

Studer, B. and Minot, M. (2018) 'Rapport d'Information No. 610'. Report submitted to the French National Assembly by the Commission des Affaires Culturelles et de l'Education. Online. www.assemblee-nationale.fr/15/pdf/rap-info/i0610.pdf (accessed 10 June 2018).

Thomas, F. (2017) 'Le baccalauréat 2017: Session de juin'. *Note d'information* No. 17.18. Paris: Ministère de l'éducation nationale, Direction de l'Evaluation de la Prospective et de la Performance (DEPP). Online. https://goo.gl/89Lc6S (accessed 10 June 2018).

Chapter 8

Standard setting in Georgia: The Unified National Examinations

Natia Andguladze and Iwa Mindadze

There is a growing debate around examination standard setting methodologies in many countries. Standard setting is a procedure of classifying examination results in several performance levels. However, not all countries set achievement standards in their examinations. To examine the wider contextual factors that contribute to the absence of standard setting in examinations, we use Georgia's Unified National Examinations (UNE), where admission examinations are cohort-referenced and the pass score is set just above what an applicant would have scored by guessing multiple choice examination item responses randomly. The UNE were introduced to combat corruption in university admissions, and the examinations have served this purpose. But studies indicate that a large proportion of students entering academic programmes are not university ready. A logical response to the issue would be setting minimum entry performance levels. There is, however, no discussion around minimum qualifications for university readiness. We argue that the absence of a debate is largely an effect of the current cost-sharing financing arrangement in the higher education system. The state is unable to financially sustain higher educational institutions, and universities have to compensate for the lack of public funding through increasing admissions numbers. The current university admissions system is a compromise balancing the interests of students, universities and the state against the growing demand for higher education and the country's inability to provide quality education. Introducing minimum standards in UNE, without reforming the university financing system or raising the quality of teaching in schools and alternative educational opportunities, would negatively affect universities' financial stability and increase the share of youth outside education and employment.

Introduction

Located between Western Asia and Eastern Europe, Georgia shares its borders with Russia to the north, Armenia, Azerbaijan and Turkey to the south and the Black Sea to the west. Its *de jure* territory is 69,700 square kilometres. The country's population was approximately 3.7 million as of 2015, with over 1.8 million living in the capital city Tbilisi. Eighty-four per cent of the population is ethnic Georgian. Other major ethnic groups include Abkhazians, Ossetians, Armenians, Azerbaijanis, Russians, Kurds and Greeks. Over 300,000 citizens are displaced from Abkhazia and South Ossetia, that is, Georgian territories that the Russian Federation occupies.

There are 2,320 schools in Georgia, with 506,659 students in public schools and 52,756 students in 236 private schools. General education is offered at three levels: primary education (six years), basic education (three years) and secondary education (three years). Primary school is the first part of the nine-year compulsory education. Students normally start at the age of six. There is a basic curriculum for each of the six primary classes. Students are taught Georgian language and literature, mathematics, history, natural sciences, arts, ICT, civic security and sports. The first foreign language must be introduced no later than the third grade. Transition to the next grade is automatic under the condition of regular attendance and a positive evaluation from the teacher. Basic education (7th–9th grades, 12 to 14+ years old) is the second stage of compulsory schooling. The school programme includes Georgian language, mathematics, history, geography, civic education, physics, chemistry, biology, arts, ICT, civic security and sports. The second foreign language is introduced in the 7th grade. Once compulsory basic education is completed, students can either continue onto secondary education, enter the first, second or third levels of professional education (UNESCO International Standard Classification of Education (ISCED) 2011 level 3 vocational stream without direct access to ISCED 2012 level 6 programmes), or leave the education system altogether. Secondary education covers grades 10 through 12. Typically, students enter at the age of 15. Attending secondary school is voluntary. The Constitution guarantees free-of-charge access to primary, basic and secondary education.

At the completion of secondary education (level 3 of the ISCED or ISCED 3), students take national school leaving examinations. Examinations are also set at the entry of the secondary vocational education stream (ISCED 3), post-secondary education stream (ISCED 4) and higher education (ISCED 6). None of these examinations are school-based, and the examinations are used solely for the purpose of judging student

competency. Current school accountability mechanisms do not use student performance in any of the above-mentioned examinations as school and/or teacher performance indicators.

This chapter is limited to issues related to the UNE. The examination, as with all other state commissioned examinations in the education field, is run by the National Assessment and Examinations Centre (NAEC). NAEC is an independent legal entity of public law under the Ministry of Education and Science (MoES) of Georgia. The Minister of Education and Science and the Prime Minister appoint the NAEC director. NAEC is financed by and accountable to the Ministry. NAEC is also responsible for the development and administration of school leaving examinations, vocational programme entry examinations, graduate programme (ISCED 7) examinations, teacher subject matter examinations, the administration of international studies (PIRLS, TIMSS, PISA, etc.) and national assessments in education.

Participation in the UNE is a prerequisite for entry into ISCED level 6 programmes in state authorized higher educational institutions. Students with secondary education are eligible for UNE examinations. Based on their performance in the examinations, students are enrolled in university programmes and are awarded state merit-based grants. Among eligible candidates, needs based grants are also awarded based on student performance in the examinations because students eligible for needs based grants are ranked by their UNE scores and then the best achievers receive the grant. Approximately 50,000 applicants participate in the examinations every year, and about 40 per cent of these applicants gain entry into university programmes. In 2014, the enrolment rate was estimated at 39 per cent.

Table 8.1: Gross enrolment rates in education (%)

Level of education	2008–2009	2009–2010	2010–2011	2011–2012	2012–2013	2013–2014
Primary (ISCED 1)[1]	104	103	103	105	104	102
Basic (ISCED 2)[1]	97	99	101	99	101	102
Secondary (ISCED 3)[1]	83	91	84	75	74	79
Tertiary (ISCED 6)[2]	26	29	31	29	35	39

Source: [1] Centre for Education Management Information System, 2014; [2] World Bank data bank

Note: Gross enrolment rate is the total enrolment in a specific level of education, regardless of age, expressed as a percentage of the eligible official school-age population corresponding to the same level of education in a given school year.

The main reason for introducing the UNE was the elimination of corruption in university admissions. Corruption in Georgia has been a longstanding issue, permeating all areas of social and economic life. As a Soviet Republic, Georgia stood out for its high level of corruption. In the 1970s, coverage of corruption in the Soviet press indicated that all areas of administration were affected (Law, 1974). Along with health services, the judiciary and housing, there were documented cases of corruption in educational institutions. In 1973, the Soviet press announced the removal of the Rector of the Tbilisi Medical Institute from his post for 'extremely flagrant violations of socialist legality and criminal actions', including manipulating the entrance examinations at the Institute 'to the benefit of his own pocket' (Law, 1974: 101). This was far from being an isolated case. According to a personal account, to enter an institute of higher education, 'payments to the "right people" were absolutely necessary' (Levy, 2007: 428).

The problem persisted after the break-up of the Soviet Union: 'bribes ranged from US$8,000 to US$30,000, depending on the prestige of the programme, according to a 2004 survey' (Rostiashvili, 2004). While most students paid bribes to tutors who served on university examination boards, politicians would trade their political support directly with the rector to gain university entrance for their family members. Only outstanding students would be able to gain admission based on their performance. For poor students and students from outside regional centres, the system did not provide much of a chance at success (World Bank, 2012).

Unsurprisingly, the primary objective of the education reform initiated in 2004 was to combat corruption in the education system. Work on a new admissions system started long before the reform, under a World Bank financed, large-scale reform preparation project. The first Unified Admission Examinations were introduced in 2005 to eliminate corruption in the university admissions process. Prior to this reform, universities had full freedom to decide on the number of students to enrol and the procedures for enrolment. The government of Georgia centralized the admissions process and linked it to achievement on standardized examinations. Because the government also changed the university-financing system, the new examination system was used to identify merit-based grant recipients.

NAEC prepares and administers examinations based on the examination framework that the MoES approves. The framework design rests on the premise that the examinations are to assess an applicant's

ability to succeed in university studies. However, the primary objective of the examinations is to function as an objective selection tool for university admission and student grant allocation. At the undergraduate level, the system does not allow universities to make decisions on an individual applicant's admission; at the application stage, applicants can choose 20 different programmes and list them by preference. After UNEs are administered, applicants are ranked within the programmes of their choice by their examination scores. Ministry officials cannot award grants (full or partial tuition waivers) to individual students either. Admitted students are awarded grants based on their UNE scores and the programme of choice.

In 2005, all applicants had to take three mandatory examinations in: the Georgian language, a foreign language and a general aptitude test. In time, other examinations were added so that university programmes could ask for an additional, fourth, subject-specific examination. For example, medical schools require an additional examination in chemistry, while economics programmes require an examination in mathematics. Also, some universities give applicants choice among various field-specific examinations. For example, to apply to a programme in medicine, students can take an examination in chemistry, physics or biology. Field-specific examinations are offered in literature, civic education, history, geography, arts, mathematics, chemistry, physics and biology. Universities usually require one field-specific examination in addition to the three mandatory examinations. The Ministry sets the minimum threshold on each examination. However, every university programme decides on the weight assigned to each examination and the additional fourth examination for its programmes. Universities can also set so-called 'minimum competency' requirements above the minimum threshold.

Entry requirements are different for applicants who finished school in ethnic minority language schools (Azerbaijani and Armenian) as well as students taking examinations in the Ossetian and Abkhazian languages. These students are required to take only one examination, the general aptitude test. The examination is offered in three ethnic minority languages (Russian, Azerbaijani and Armenian). Some university programmes (e.g. sports, arts, music) administer additional, university-based examinations. All UNE examinations are offered in Georgian and Russian.

The assessment process

NAEC subject matter units develop the UNE examinations based on the content standards provided in the national curriculum. The subject

matter groups also consult with teachers and subject matter experts to develop test specifications. The review team is provided with statistical analysis of the previous year's UNE examinations. Test items (and open-ended item rubrics) are reviewed by subject matter experts and teachers to ensure that they conform to the examination framework (e.g. evaluation of content validity to ensure that subject related skills and knowledge are appropriately covered). Some of the items are pre-tested with a group of volunteer applicants to assess item difficulty, discrimination and gender differential item functioning. Volunteer applicants are selected both from urban and rural schools so that the pilot group is representative of the UNE applicant population. Over 1,000 volunteers participate in the pre-test in every examination each year. The numbers vary for each examination. The NAEC research team runs pre-test statistical analysis and provides the information to the subject matter teams. Based on the item pre-test statistics, subject matter teams choose or modify the pool of examination items.

All examinations include a mix of open- and closed-ended items of low, medium and high difficulty (see sample items in Annex 1). Short answer, essay and computational items are used in open-ended items. Closed-ended items are usually true or false, matching or multiple choice. Each examination has three or more versions. All examinations are cohort-referenced in order to identify the best achieving students to award merit-based grants. The examinations programme is posted online a year prior to the examinations. The programme describes the content of the examinations and the skills that applicants should demonstrate.

Examinations are administered once a year in 14 centres around Georgia. Entrance examinations are printed abroad. The rationale behind using this arrangement is to ensure that examination content remains confidential. The sealed examinations are sent back to Georgia and delivered in police cars to the vaults of the National Bank, where they are stored until examination day. Some 700 local proctors undergo training to monitor the examinations. Examinations are identified by barcode rather than student name to help eliminate bias during marking. Closed-circuit cameras are installed in every examination room. This was originally done to detect and prevent illicit practices by students and proctors but later on became a tool for parents to monitor the process from a waiting room outside.

Since 2008 NAEC has used eMarking to score open-ended items. It is a blind, double-marking process. Each item is marked by two scorers who

work independently from each other. Expert scorers from NAEC subject matter teams monitor scorer agreement. Scorers are usually subject matter teachers or university professors who are trained by the NAEC subject matter teams. Each year, there is approximately 30 per cent rotation in the scoring team.

Applicants can access their examination results using unique login information they are provided with. Using that, they can access their examination papers, scoring guides and their scores on each of their examination assignments. After getting access to their marked examination papers, applicants have two weeks to appeal.

After scoring is complete, raw scores are converted into scaled scores. Because there are usually multiple versions of one examination (e.g. English language), scores across different versions of the same examination are first equated using percentile rankings. Scores are then standardized to make different subject examinations comparable using the mean scores of each subject examination. Passing scores in all examinations are set just above the score an applicant would obtain by guessing closed-ended question responses randomly.

Examination predictive validity studies are conducted in two-year rounds. The most recent validity study shows that the strength of the relationship between student Grade Point Average (GPA) and their performance in university examinations varies by university. The relationship is higher in more prestigious private universities, and lower in the least prestigious universities that enrol students with the lowest performance on the examinations. For example, in one of the country's most selective universities, the correlation is stronger between Georgian language and GPA ($r = 0.38$) than with the general aptitude test ($r = 0.30$) or English examination ($r = 0.25$). Also, the relationship between GPA and examination scores is more pronounced during the first semester of studies and decreases over time (NAEC, 2009).

Assessment approaches and their social effects and implications

The introduction of the UNE has arguably met its objective to eliminate corruption. It has also been claimed that the UNE has improved access to higher education for students from outside the capital (World Bank, 2012). A decade after the first round of examinations, the wider public as well as the school community still trust the UNE. According to a 2013 survey (CRRC, 2013), the majority of Georgians, when asked about 'the best way

of organizing admissions to university in Georgia', chose unified admissions to universities (see Table 8.4 in Annex 2). The majority of teachers, school principals and parents believe that the UNE were a 'very successful' or 'successful' reform (see Table 8.5 in Annex 2).

There is, however, growing concern over some characteristics of the Georgian education system that are indirectly linked to the university admission system, specifically the methodology of identifying the passing score in the UNE. Recent labour market studies point towards a skills mismatch resulting in a high unemployment rate among Georgian youth. The skills mismatch has been partly explained by the quality and relevance of vocational and tertiary education programmes. But, the studies also point towards an oversupply of higher education graduates (Bartlett, 2013; Bardak, 2011; World Bank, 2013).

Each year, about 70 per cent of secondary school graduates apply to universities, and an increasing number of these students are enrolled in undergraduate academic programmes. Not all students, however, are academically prepared for university programmes. As a number of international assessments have shown, a large share of Georgian students reaches the secondary level (ISCED 3) without basic reading and mathematics skills. For example, the Programme in International Student Assessment 2015 results show that over half of the 15-year-old population in Georgian schools perform below the baseline level (level 2) in reading, 'at which students begin to demonstrate the reading skills that will enable them to participate effectively and productively in life' (OECD, 2016: 164). These students could be considered functionally illiterate. Twenty-five per cent perform at the basic proficiency level, and the remaining 23 per cent perform above the baseline level. Reasonably, many Georgian students who perform below the baseline level in reading will find it very challenging to study in a university.

There is a growing understanding that enrolment in academic programmes is an issue. The Minister of Education and Science has recently announced the ministry's plan to raise the passing score for the UNE. The underlying rationale behind the plan is to raise university entry standards. However, in the absence of performance standards in the examinations, it remains unclear what the new minimum passing score would mean for students and universities. It could be argued that the existing pass score setting approach creates a communication problem among the parties involved. We would argue that using a standard setting methodology in the pass score identification process would give universities the opportunity

to make better informed decisions on passing scores for enrolment in their programmes. It would also give schools information on the gaps in their students' knowledge and provide students with a better understanding of their readiness for university examinations.

It is worth noting that using standard setting in examinations is an established practice in Georgia. NAEC uses standard setting in teacher certification examinations and national assessments. Standard setting in certification examinations is based on the Angoff method. In the National Assessments of Educational Achievement, the bookmark standard setting method is used. The experience could be applied to UNE. However, larger contextual factors would greatly hinder the possibility of such a change in the UNE. We argue that the existing pass score identifying approach is tightly linked to other system characteristics. Transforming UNE examinations requires reforming (1) university financing schemes, (2) university accountability mechanisms and (3) technical and vocational education programmes. Understanding these three aspects of the education system in Georgia sheds light on the challenges that the system could face if UNE moves to a standard setting model.

University financing

Public spending on tertiary education is very low, and tertiary education has relied heavily on student tuition fees which come from students' households. Approximately 1.2 per cent of the total government budget is allocated to tertiary education, which is significantly less than in most developed countries, including those of the former Soviet and Communist bloc (e.g. 2 per cent in Estonia and 2.4 per cent in Poland (OECD, 2016)).

Research funding is also low in Georgia. Public and private spending on research as a share of GDP in Georgia (0.2 per cent) is well below the average for middle-income countries (0.6 per cent), the Commonwealth of Independent States (0.4 per cent) and Central and Eastern European countries (0.9 per cent). Because income generation capacity is low and philanthropic funding is rare, Georgian universities are largely dependent on student tuition fees. Seventy per cent of public university sector revenues and almost 100 per cent of private sector university revenues come from student tuitions. Even the two largest research universities generate over half of their revenues from student tuition fees.

This financing arrangement is a result of a cost-sharing policy. The government promoted and implemented the policy in 2004 together with other neo-libertarian reforms in the field of education (e.g. school choice

and the 'money follows student' financing modality in secondary education) following the Rose Revolution. These reforms have shaped the current education landscape in the country.

With little to no other funding sources besides individual contributions, universities are highly dependent on enrolment numbers. Therefore, the government has gradually allowed universities, both public and private, to increase the enrolment quotas. From 2005 to 2011, the number of available seats in academic programmes increased by 120 per cent. As a result, the admissions rate increased from 53 per cent in 2005 to 77 per cent in 2011. The enrolment rate increased from 26 per cent in 2009 to 39 per cent in 2014.

Table 8.2: TE application and admission statistics in 2005–2012, academic programmes only

Year	Available seats	Number of applicants / number of available seats	Admissions rate (%)
2005	17,501	1.8	53
2006	19,714	1.7	59
2007	15,501	2.5	49
2008	15,779	1.5	76
2009	25,054	1.2	80
2010	33,681	1.1	70
2011	33,988	1.0	77
2012	38,738	0.9	76

Source: The author's calculations based on the data base provided by the National Assessment and Examination Centre

University accountability

The current university accountability system is also a piece of the puzzle. All the accountability mechanisms developed since 2005 focus on inputs. Currently, Institutional Authorization (IA) determines the enrolment quota. IA is basically a mechanism to regulate enrolment rates. Programme accreditation awards the right to introduce/maintain undergraduate or graduate programmes and is implemented through a peer review of education programme proposals. Peer evaluation is also used to periodically assess the

quality of programmes through face-to-face interviews with students and university professors.

Other accountability mechanisms are absent from the system, and information on the quality of education processes is not available to the wider public. There is not a single piece of information about the processes, student engagement, student learning outcomes, or any quality indicator available to the public except for the number of applicants per available seat or the mean examination scores of applicants.

Moreover, unlike arrangements in many other countries, universities are not held accountable for the quality of teaching or research. Since universities are not held accountable for their outcomes, including for what their students learn, no one can truly judge the quality of their programmes. Because outcomes do not matter, universities as institutions are not concerned with the university readiness of their students. Most universities set low entry barriers by assigning lower weights to more challenging examinations. In 2016 university admissions, of 74 public and private universities, only five universities set a so-called 'competency limit' above the default minimum threshold in one or more examination. For example, Tbilisi State University, the largest and oldest research university in the country, set its threshold at 40 per cent of the maximum score in every required examination for all programmes. Yet the requirements are not linked to standards.

Other post-secondary educational opportunities

As the discussion above shows, many students who aspire to study in post-secondary academic programmes are not ready for them. The newly appointed Minister of Education and Science has raised the issue of increasing university entry requirements and has devoted lengthy public speeches at universities and on social media to the subject, addressing students about the importance of making good career choices based on personal aspirations and disregarding social pressure. These students need to have alternative, attractive and less academically challenging post-secondary educational opportunities. However, these opportunities are limited. After the break-up of the Soviet Union, the system of vocational educational institutions collapsed. Technical colleges were left with very little funding, resulting in the deterioration of the quality of teaching staff, equipment and infrastructure. Currently, ISCED 3 and ISCED 4 level programmes can accommodate only about 15 per cent of secondary school graduates (see Table 8.3 below). Raising the bar on UNE would result in a further increase

in the share of youth not in employment, education and training (NEET) which is already very high compared to all EU countries. According to 2013 National Household Survey data, Georgia registers a NEET rate of 31 per cent for the 15–24 age group, which is 18 per cent higher than the EU average and 11 per cent higher than Bulgaria, the country with the highest NEET rate among EU countries (Bardak *et al*., 2015).

Table 8.3: Capacity of ISCED 3 and 4 (2011) Level State Providers in 2010

	Georgia	Tbilisi	Regions
Number of state colleges	20	7	13
Number of available seats	7362	2638	4724
Share of available state funded places (%)	62.5	58.2	64.9
Number of applicants	7385	3719	3666
Number of enrolled students	5042	2387	2655
Share of state funded students (%)	63.4	56.8	69.3

Source: MoES, 2011

Discussion

In many countries, setting performance levels is a part of the examination process and has strong implications for students, and in some education systems, for schools and educators as well. Therefore, the methodologies used in setting performance levels have drawn increasing attention from the education and research communities.

Internationally, not all examinations use performance levels to define cut-off scores. One such example is the UNE in Georgia that has been used in university admissions since 2005. The cut-off scores in the examinations are set just above the score an applicant would obtain by guessing closed-ended question responses randomly. Judging from the distribution of students by proficiency levels in national and international assessments, many of the students who enter higher educational institutions do not have the literacy and numeracy skills needed to succeed in university studies. Since the UNE's introduction, the education community has remained comfortably numb about the absence of standard setting in what is arguably the country's most widely covered and discussed set of examinations. We argue that an important determinant of the country's

choice of the examination model is the current set of financing and accountability policies and practices.

There is very little research on how examinations are related to wider contextual factors. Noah and Eckstein (1989) reviewed examination systems in eight countries and identified contextual and wider policy characteristics that define the countries' examination systems. The authors claimed that the characteristics of examination policies and practices represent trade-offs among competing values and 'while seeking to increase perceived benefits in one direction, a nation almost inevitably gives up some benefit or exacerbates some problem in another direction' (Noah and Eckstein, 1989: 17).

Noah and Eckstein used the United States as an example of rejecting traditional extended-answer type examinations due to the high financial and logistical burden in the face of a growing number of applicants. Thus, the trade-off was made between validity on the one hand and accessibility and objectivity of examination systems on the other hand. In 1976, Japan introduced a two-stage examination system – examinations administered both at schools and by universities – giving higher educational institutions more control over the make-up of their student population. The trade-off of the change has been a high cost for the families of candidates as the additional examinations have led to thousands of dollars in post-school preparation costs and travel costs associated with going to distant cities to sit for the second-level examinations. In the late 1980s, France diversified its unified school leaving examination (the baccalauréat) into a complex examination system to accommodate the growing variability in competencies of school graduates. From a single nationally comparable examination administered to all candidates, the baccalauréat was transformed into an examination system with a strongly demarcated hierarchy of prestige with mathematical options at the top and vocational options at the bottom of the hierarchy. The cost of the diversification has been the loss of comparability across candidates, a problem that, as Noah and Eckstein claimed, further exacerbated the devaluation of Baccalauréat due to the devolution of examination administration responsibilities to regional academies (Noah and Eckstein, 1989).

Georgia provides an interesting case for examining how wider systemic characteristics determine the design of examinations and specifically the absence of standard setting from examinations. In Georgia, there are some wider contextual factors that are very different from those of other countries. The most important function of the university examinations is

to keep admissions and grant allocation free from corruption. Also, unlike many countries with high stakes examinations in place, Georgia does not use examinations for accountability purposes, for example to evaluate school or teacher performance.

Two distinctive features specifically relevant to examination standards are that the largest share of university revenue comes from student tuitions, and universities are not held accountable for the quality of educational processes or student outcomes. So, even if universities enrol students who are going to fail, there are no mechanisms for holding universities accountable for student failure. The combination of these two factors provides incentives for a majority of universities to enrol as many students as they can irrespective of the students' readiness for university study.

Setting minimum admissions competency in terms of standards would be a very challenging task considering the skills of the students at the bottom of the distribution. We argue that defining minimum competencies for university readiness or acquisition of the competencies covered in the school curriculum would exclude many students who would otherwise find a place in universities. This would translate into decreased enrolment numbers in many, particularly public, universities in the regions and the traditionally least popular programmes such as education and the sciences. Universities would lose a considerable share of their revenue and would be forced to close some programmes. Moreover, if some of the students who currently study in universities were rejected, they would be forced to look for admission at ISCED 4 and ISCED 3 level programmes, which do not have the capacity to accommodate them. Thus, if setting standards in admission examinations raises the bar, *ceteris paribus*, then the proportion of youths outside the labour market, education and training, which is already high by any standard, will rise more.

The cost-sharing policy in higher education has had its price. In Georgia the price seems to be the quality of education which has many manifestations, deliberate and unintended, such as the absence of performance related accountability instruments in schools and universities. The absence of minimum standards in university admissions seems to be the compromise that the government and education community have reached, leading to a quiet equilibrium which balances the interests of students, universities and the state against a background of growing demand for higher education and the country's inability to provide quality education.

Annex 1. Sample items from the UNE examinations

Problem 1 **1 point**

Which number listed below belongs to the interval $(0,7; 0,8)$?

a) $\frac{3}{5}$ b) $\frac{7}{9}$ c) $\frac{6}{7}$ d) $\frac{8}{9}$

Problem 8 **1 point**

A car moves at a constant speed from Tbilisi to Kutaisi. By eight o'clock in the morning the car covered $\frac{1}{6}$ part of the planned route, and by 11 o'clock in the morning of the same day $\frac{8}{9}$ part. What part of the planned route did the car cover by 10 o'clock and 30 minutes in the morning of the same day?

a) $\frac{65}{108}$ b) $\frac{57}{108}$ c) $\frac{8}{18}$ d) $\frac{83}{108}$

Problem 31 **2 points**

Solve the system of equations

$$\begin{cases} \frac{3}{2}x + 2y = 7 \\ 2x - 3y = 5 \end{cases}$$

Problem 32 **2 points**

Two business partners have divided the profit in the amount of 80500 GEL in the ratio of 2:5. How much money did each one get?

Problem 35 **3 points**

The total number of white and black balls in a box is 42. If one ball is drawn out of the box at random, what is the probability that this ball is white if it is known that adding 6 new white balls in the box will cause the increase of this probability by $\frac{5}{4}$-times?

Problem 39 **4 points**

Several workers did their work in 14 days. If there had been 4 workers more and a working day – 1 hour longer, the same job would have been completed within 10 days. If there had been 10 workers more and the working day – 2 hour longer, the same job would have been completed within 7 days. How many workers worked and what was the duration of the working day in hours, if it is known that labor productivity of all workers is the same?

Figure 8.1: Sample items from 2016 Mathematics examinations (a full version of the examination is available from: www1.naec.ge/images/doc/EXAMS/math_2016_final_eng.pdf)

Task 1: Listen to ten texts (1-10). For each of them answer the question given. Mark the correct choice (A-D). You have 20 seconds to look through the task. You will hear each recording twice.

1. Where is the dialogue taking place?

A. At the hotel

B. At the airport

C. At the café

D. At the theatre

Task 3: Read the text. Then read the statements which follow and decide whether they are True (T) or False (F).

Shakespeare Folio found in France

A rare and valuable Shakespeare First Folio, regarded as the most important book in English literature, was discovered in a public library of Saint-Omer, a small town in northern France. The Folio was discovered at the end of 2014 by a librarian and medieval* literature expert Rémy Cordonnier, while he was preparing an exhibition of the historic links between France and England. 'The First Folio' is the name of the collection of Shakespeare's plays prepared and printed by his friends and colleagues John Heminges and Henry Condell in 1623, seven years after Shakespeare's death. Some of Shakespeare's plays had been published before 1623 too, but the First Folio is considered to be a book of great significance because the text of the plays in it are thought to be the most reliable ones, not modified by anybody. The First Folio collects 36 of Shakespeare's 38 known plays for the first time.

Experts believe that originally 800 copies of the Folio were produced, though most of them have been lost. The Folio, found in the French library, was the 233rd known surviving one. The rest are kept in different museums and private collections of the world. France owns only two copies, including the one discovered in the library. Each discovery of the Folio attracts the interest of literature experts as well as collectors. One of the Shakespeare First Folios discovered in recent years was sold at Sotheby's auction in 2006 for 5.2 million dollars.

The copy of the First Folio discovered in the French library was heavily damaged; it didn't have several introductory pages or the title page. This may have been the reason why the book lay in the library for almost 200 years and nobody was able to identify its significance - that it was the famous collection of Shakespeare's plays published in 1623. One of the world's most known Shakespeare experts, professor Eric Rasmussen from the University of Nevada, USA, was the first to study the texts of the Folio. He found out that it contained several handwritten notes, which may make it clear how the plays were performed in Shakespeare's time. For example, in one scene from the play *Henry IV*, the word 'hostess' was changed to 'host' - possibly reflecting the fact that in early performances only men acted, so a female character was turned into a male.

8

However, the Folio is not the rarest book the Saint-Omer library owns. It also has a Gutenberg Bible, of which fewer than 50 copies have survived. The Gutenberg Bible is the first printed bible dating back to 1450 and is known as a starting point of the printed book in Europe. The information about the Shakespeare First Folio, as well as any other information related to the great playwright, is even more valuable this year, when Britain celebrates the 400th anniversary of Shakespeare's death.

*medieval - შუასაუკუნეების

True (T) or False (F)?

1. The person who discovered the Folio was an expert in medieval literature.
2. The First Folio is the name of the book prepared by Shakespeare himself.
3. The First Folio contains all 38 plays written by William Shakespeare.
4. Most of the original Folios have been lost.
5. The Folio, discovered in the library, is the second copy existing in France.
6. None of the pages in the newly discovered Folio were torn off or lost.
7. The notes made on the texts may clarify some facts about how plays were performed then.
8. The Gutenberg Bible is older than the Shakespeare First Folio.
9. A very important date is celebrated for British literature lovers this year.
10. The text is about several new discoveries in the French library.

9

Task 8: Read the text and put the verbs in brackets in the correct form. Do not copy the extra words from the text on the answer sheet.

Dear Helen,

I hope you are well. You know my friend, James. He is a nice guy but I ….. (1. have) some problems with him recently. One day he didn't answer my calls and the next day he came late and didn't tell me where he ….. (2. be). But last weekend was a real nightmare. It was Saturday morning and James told me he ….. (3. pick) me up at 8 pm to go out for dinner. It was already nine o'clock but James ….. (4. not/appear) yet. I ….. (5. begin) to get worried when my telephone rang. A stranger told me that James ….. (6. arrest) but did not tell me why. I hurried to the police station. When I got there, a detective ….. (7. question) him. I waited until the interrogation finished. But even after that I ….. (8. not/allow) to talk to him. I ….. (9. tell) that he had been arrested because he had hit a dog with a car and killed it. They told me that James had to pay a fine. Unfortunately I didn't have any money with me. So, he had to spend the night at the police station.

It was already very late, so I couldn't bother any of my friends. On Sunday morning I ….. (10. rush) to one of my friends, Susie. But as Susie ….. (11. spend) that weekend with her parents in the summer house, I had to take a train. Luckily Susie was able to lend me money and I released James from the police station.

I promise, next time if I have happier things to tell you, I ….. (12. let) you know immediately.

Best wishes,
Nelly

19

Task 9: The advertisement given below is taken from an online newspaper. Read the advertisement and write an email to New York School of Business asking for more information about the details which are indicated. The beginning is given on the answer sheet. Do not write your or anybody else's name or surname in the letter.

Are you interested in Business Administration? If so, read this advert carefully.

New York School of Business offers a training course in Business Administration. The school is located in the very **centre** of New York. The course offers classes in **several** disciplines. Participants can take intensive English language classes **in the evening**. The course lasts for three months. Contact us at ba@gmail.com

Where exactly?

When exactly?

How many?

Task 10: Read the essay task and write between 120-150 words.

Some people think that schoolchildren should spend more time on sports activities, such as, class-to-class or school-to-school sports competitions. Do you agree or disagree with this opinion? State your opinion and support it with reasons and examples.

Figure 8.2: Sample items from 2016 English language examinations (a full version of the examination is available from: www1.naec.ge/images/doc/EXAMS/eng.abit.%201.%202016.pdf)

Annex 2. Additional tables

Table 8.4: Public opinion on 'What is the best way to organize admissions to the universities in Georgia?'

Frequency distribution (%)	
Unified admissions managed by a centralized body and based on standardized exams	47
Admission managed by universities based on standardized exams	14
Universities managing both exams and admission	15
Other	0
DK/RA	24

Source: CRRC, 2013. Retrieved from http://caucasusbarometer.org/en/cb2013ge/UNIVADM/ (accessed 11 August 2016)

Table 8.5: School community's attitude about the degree of the success of UNE reform intervention

How would you rate the degree of success of the reforms (in %)?	**School principals**	**Parents**	**Teachers**
	(n = 165)	(n = 3237)	(n = 194)
Very successful	51.3	24.5	31.4
Successful	41.5	57.3	55.8
Neither successful nor unsuccessful	2.3	5.2	5.8
Unsuccessful	0.4	3.6	4.6
Very unsuccessful	0.4	2.1	1.1
Don't know	4.1	7.3	1.4

Source: NAEC, 2015

References

Bartlett, W. (2013) *Skill Mismatch, Education Systems, and Labour Markets in EU Neighbourhood Policy Countries* (SEARCH Working Paper WP5/20). Barcelona: SEARCH.

Bardak, U. (ed.) (2011) *Labour Markets and Employability: Trends and challenges in Armenia, Azerbaijan, Belarus, Georgia, Republic of Moldova and Ukraine*. Turin: European Training Foundation.

Bardak, U., Maseda, M.R. and Rosso, F. (2015) *Young People Not in Employment, Education or Training (NEET): An overview in ETF partner countries*. Turin: European Training Foundation. Online. www.etf.europa.eu/webatt.nsf/0/BFEEBA10DD412271C1257EED0035457E/$file/NEETs.pdf (accessed 2 June 2017).

CRRC (Caucasus Resource Research Center) (2013) Caucasus Barometer Annual Household Survey. Regional Database of 2013. Online. http://caucasusbarometer.org/en/downloads/ (accessed 28 July 2018).

EMIS (Centre for Education Management Information System). *Gross Enrolment Rates in General Education 2008–2014*. Tbilisi: EMIS.

Law, D. (1974) 'Corruption in Georgia, critique', *Journal of Socialist Theory*, 3 (1), 99–107.

Levy, D. (2007) 'Price adjustment under the table: Evidence on efficiency-enhancing corruption'. *European Journal of Political Economy*, 23 (2), 423–47.

MoES (Ministry of Education and Science). (2011). Vocational Education in Facts and Figures, 2011. Tbilisi: MoES

NAEC (National Assessment and Examination Center) (2009) The National University Admission Examination in 2009. Tbilisi: NAEC.

NAEC (National Assessment and Examination Center) (2015) National Assessment in Mathematics in the 9th grade, 2015. Dataset. Tbilisi: NAEC.

Noah, H.J. and Eckstein, M.A. (1989) 'Tradeoffs in examination policies: An international comparative perspective'. *Oxford Review of Education*, 15 (1), 17–27.

OECD (Organisation for Economic Co-operation and Development) (2016) *Education at a Glance 2016: OECD indicators*. Paris: OECD Publishing.

Rostiashvili, K. (2004) *Corruption in the Higher Education System of Georgia*. Tbilisi: Transnational Crime and Corruption Center, Georgia Office.

World Bank (2012) *Fighting Corruption in Public Services: Chronicling Georgia's reforms*. Washington, DC: World Bank.

World Bank (2013) *Georgia: Skills mismatch and unemployment labor market challenges* (Report No. 72824-GE). Washington, DC: World Bank. Online. http://documents.worldbank.org/curated/en/999371468242985088/pdf/728240ESW0Geor00Box377374B00PUBLIC0.pdf (accessed 30 May 2018).

World Bank Data bank. *School enrollment, tertiary (% gross)*. Retrieved from https://data.worldbank.org/indicator/SE.TER.ENRR?locations=GE (accessed 23 July 2018).

Are low standards the same as no standards?

Steven Bakker

Since the Rose Revolution in 2003, Georgia has made important steps in banishing the pervasive corruption it inherited from the time it was a Socialist Soviet Republic. Major accomplishments that are still recognized by the public at large are the reorganization of the police force and replacing the many university-run entry tests with the Unified National Examination (UNE). The chapter written by Andguladze and Mindadze illustrates the necessity of this measure by the time of its introduction in 2005, but argues that the UNE is not used as a standards-based hurdle to separate those who would be fit for academic studies from those who are not, a function it should in fact have.

The UNE is a professionally set and administered large-scale high-stakes test, the quality of which in all its aspects can easily compete with similar tests in countries with much longer experience in public examinations. The UNE definitely has an in-built standard: its test components represent the knowledge and skills field experts believe a candidate eligible for university studies should be able to demonstrate. Such a standard, not supported or operationalized by validated standard setting methods, is not uncommon. Certainly in countries with a long tradition in administering national exams, standards are communicated by agreed pre-defined cut scores that aim at keeping the percentage pass scores the same from one year to another. This approach is justified by the assumption that the overall abilities from different cohorts do not differ significantly, and exams may be set with the pre-defined cut scores in mind, if they do not differ significantly in difficulty grade, from one year to the other. The origin of such cut scores, for example 50 per cent of the maximum score, is buried somewhere in history but nobody seems to bother too much about where they came from. Yet another example is the standardized admission tests such as the US SAT that does not come with any cut score at all. Users set their own, based on what has proven over time to be the minimum score needed for being a successful student at their institution.

UNE exams are standardized: a test matrix is implemented, the difficulty grade is kept constant over the years using pre-test data, and a conscious decision underlies the minimum score for passing. The problem

the authors address, though, is the fact that the minimum scores decided by the Ministry of Education and most individual universities do not reflect the minimum competence that experts in charge of setting the tests believe students should demonstrate in order to be eligible for academic studies; rather, they are kept deliberately low to avoid high failure rates. The chapter written by Andguladze and Mindadze shows that this is in the interest of most stakeholders: the Ministry is not stuck with a sizable number of school leavers that have no place to go, students have easy access to tertiary education, and universities keep financially afloat. Critics will maintain that this is the price of low-quality education and point to the over-representation of students at the lowest levels of the National Assessments of Educational Achievement (set using a traditional, validated standard setting method). Others may reason that it is only fair for the cut scores to keep step with the low level of knowledge and skills that currently emerge from Georgian public education. Following the intended gradual improvement of the quality of the educational system, especially the quality of teachers and school leaders, they should raise over time, though, to match the academic standards applied in most modern economies. Then the moment will arrive to bring together representative panels to agree on described minimum competence levels for eligibility to academic studies, and scientifically valid methods to implement and safeguard these over time.

Social needs and standard setting

Gordon Stobart

This is an important case study of an examination which is designed to meet particular social needs: fair selection for university that encourages, because of the university funding system, mass entry into higher education. It illustrates the social determinants of any exam system and some of the trade-offs that have to be made in operating a selective examination.

The authors are to be congratulated on their thoughtful, open and critical account of the purposes and uses of university entrance examinations in Georgia. They make it explicit from the outset that the Unified National Examinations (UNE) have primarily been developed to combat corruption in university selection, corruption that was endemic in all walks of life in the Soviet era (see Bethell and Zabulionis, 2012). To this end the emphasis has been on the *integrity* and *reliability* of the system rather than on the validity with which it selects for the demands of university study.

This is in part because a key driver in the standard setting process is university funding. In a country where funding for education is comparatively low, and universities depend overwhelmingly on student fees, large numbers of fee-paying students need to pass the UNE. This is unlike many countries where universities select, rather than recruit, students; thus standard setting is used to ration the numbers qualifying for higher education. However, the principle is the same: the selection processes for higher education, as an agent of the social system, influence standard setting.

The authors are clear that the Georgian UNE is not a traditional standard setting examination based on some form of criterion-related standards. Like, for example, the American system, the grading task is to determine cut scores and to provide a normative distribution. They recognize the lack of information this provides in terms of what students know and can do. The more serious problem, however, is the low level of performance that university recruiters accept in order to fill places. This means that the system is setting cut scores 'just above the score an applicant would obtain by guessing close-ended responses randomly' (p. 144). So students are knowingly being selected for university with skill levels that mean they will not cope. From outside the system, one is tempted to ask why the exam board tolerates this. From within the system, in which vocational training

is even weaker and youth unemployment is at a comparatively high 30 per cent, there may seem little room for manoeuvre.

Returning to the integrity and reliability of UNE, Andguladze and Mindadze show how Georgia's assessment system is capable of sophisticated standard setting procedures. The dilemma is that they are not validly used in the UNE. The pressures to have as many students as possible pass means that standard setters are not making qualitative decisions to select those students who may benefit from a university education.

Paradoxically, given the quality assurance procedures in place to prevent fraud and bias, I would see the exam system as having leapfrogged, in terms of technology and quality, some of the more traditional systems, which are still trying to extricate themselves from old examination processes. Georgia has pre-testing, double-marking of anonymous scripts, high security around the development and delivery of papers and reliability checks during marking. Consideration is given to minority groups and students can access their results and appeal them. These features all point to a system determined to offer fair examinations. As a result, the UNE appears to enjoy high public and political support, a key requirement of any public examination system.

I was left curious as to why the results are not being used for any form of school accountability. This has proved irresistible for policymakers and educational authorities elsewhere. What are the social forces preventing it? This is not to argue that using results for school accountability is necessarily a social good, but what incentives do schools have to improve the performances of their students if most can qualify for university? The question also applies to higher education. We have no information on the quality of the outputs from the system. Is it more about keeping students for funding purposes than developing their learning? In what ways does the system need to change, and what is the role of the UNE, particularly its standard setting role, in this? Just as impressive steps have been taken to produce a fair examination, what steps can be taken to improve its validity with its main purpose selection for the demands of higher education, which is not happening at present?

A major contribution of this case study is to raise wider questions about the social purposes of examination systems, the compromises they involve and their integrity in meeting these demands. Standard setting is not an autonomous process. It is part of a complex social web.

Reference

Bethell, G. and Zabulionis, A. (2012) 'The evolution of high-stakes testing at the school–university interface in the former republics of the USSR'. *Assessment in Education: Principles, policy & practice*, 19 (1), 7–25.

Chapter 9

Standard setting in Ireland: The Leaving Certificate

Hugh McManus

Introduction

Ireland

Ireland is a sovereign state that covers 83 per cent of the island of Ireland, the second largest island in the British Isles and part of the continent of Europe. It has a population of 4.8 million (2016 census). It was part of the United Kingdom until the establishment of the 'Irish Free State' in 1922. Full independence came with the adoption of a new constitution in 1937, which named the state 'Ireland' (or 'Éire' in Irish), and the country was officially declared a republic in 1949. The remaining 17 per cent of the island of Ireland forms Northern Ireland, which remains a part of the United Kingdom. While Ireland's first official language is Irish, the mother tongue of the great majority of the population is English.

Ireland is a member of the European Union. It has a modern knowledge economy, relying on services and high-tech industries. Its GDP per capita consistently ranks it nominally among the wealthiest countries in the world, but its GNP is significantly lower than its GDP, due to the large number of multinationals based there.

Education system and the Leaving Certificate

The minimum school leaving age is 16, which generally coincides with the end of lower second-level education. However, there is a retention rate of over 90 per cent to the end of upper second level, and Ireland has a large and growing proportion of tertiary graduates. The proportion of the population aged 25 to 34 having a tertiary qualification is the second highest in the EU, at 52 per cent (OECD, 2016).

Upper second-level education is referred to as senior cycle, and students are typically 18 years old on completion. This cycle consists of an optional Transition Year, followed by a two-year Leaving Certificate programme, of which three variants are available. The vast majority of students follow the Established Leaving Certificate programme or the

Leaving Certificate Vocational Programme, which are almost identical, to the extent that the students of the latter programme are usually considered a subset of the former. The third available programme, the Leaving Certificate Applied, caters for about 5 per cent of the cohort. It is considerably different, and its completion does not meet the requirements for direct entry into a tertiary degree programme. For the remainder of this chapter, references to the Leaving Certificate examinations should be taken to mean those of the Leaving Certificate Established programme, including the Vocational programme, but not including the Leaving Certificate Applied.

Students typically take about seven subjects. While Irish is officially the only compulsory subject, almost all students also study English and mathematics, and over two-thirds study a third language. The remaining subjects are selected from a range of arts, science, business and applied science (including technological) subjects.

The examination dates from 1924 and was run by the Department of Education and Skills until the government established the State Examinations Commission (SEC) in 2003. The curricular programme and individual subject specifications are drawn up by the National Council for Curriculum and Assessment, usually through an extensive consultative process, following which the Minister approves them for implementation.

Given the comprehensive nature of provision and the very high retention rates, the examination must cater for a broad range of student achievement. Examinations in each subject are offered at two levels – Higher and Ordinary, with an additional Foundation level in Irish and mathematics. Schools typically have separate Higher and Ordinary level classes for English, Irish and mathematics and mixed-level classes for other subjects. In most subjects, the syllabi for the two levels differ in content, with the Ordinary being a subset of the Higher. In some cases, the content is the same, with differentiation achieved through the level of challenge of the examination papers. For example, the syllabus for the modern European languages states: 'While the syllabus is the same for both levels, the performance targets will involve language use of varying degrees of complexity' (Department of Education and Science, 1995). Up to 2016, results were issued as grades on the scale: A1, A2, B1, B2, B3, C1, C2, C3, D1, D2, D3, E, F, no grade. Results from 2017 onwards are issued on a scale from 1 (highest grade) to 8 (lowest grade) at each level.

The qualification spans levels 4 and 5 on Ireland's National Framework of Qualifications, which equate respectively to levels 3 and 4 on the European Qualifications Framework.

Use of results for tertiary entry

In addition to its primary purpose of certifying achievement on exit from second-level schooling, the Leaving Certificate also serves as a selection mechanism for entry to third level. For the great majority of courses in universities and other higher education institutions (HEIs), it is the sole basis on which entry decisions regarding school leavers are made. HEIs generally require at least two subjects to be successfully taken at Higher level to meet minimum entry requirements for an honours bachelor degree programme, but competition means that actual entry requirements for many courses are much higher than the minimum entry requirements.

Students do not apply directly to HEIs, but instead apply through the Central Applications Office, a private company established by the participating institutions for this purpose. While the individual institutions retain autonomy over their own entry criteria, they have all agreed to treat applications in the same way: for each course, there are general and subject-specific minimum entry requirements, and among all applicants who meet these criteria, places are awarded on the basis of a composite score calculated from Leaving Certificate grades. The grades are transformed into scores on a particular scale and the best six are added to give the points score. Apart from the fact that 'bonus points' have been awarded for grades in Higher Level mathematics since 2012, all subjects are equally weighted. There is therefore an inherent assumption that grades obtained in different subjects are equivalent. It may also be noted that the placement of Higher and Ordinary level grades on a common points scale establishes a de facto linkage in currency between these grades.

The points system is one of pure supply and demand. The points required for admission to a course is a function of the number of places, the number of applicants and the points 'wealth' of the applicants. The points required for entry into any course is not known until all of the relevant processing of results for all applicants is done. An increase in the number of applicants or a decrease in the number of places will increase the points cut-off score for a course. Any course that has a tendency to attract high-achieving applicants will also tend to have a higher cut-off score, but this is critically dependent on the ratio of demand to supply. If there are plenty of places available, high achieving candidates do not push others out of the market, so the points cut-off can remain low. Nevertheless, the points cut-off score for entry into a course has come to be regarded as a proxy measure of the prestige or quality of the course. This can have a vicious-cycle effect, with such courses becoming more attractive to higher-achieving students

purely because of the difficulty of getting into them. Given that HEIs are naturally interested in attracting the best students, it has been argued that this creates a perverse incentive to create increasingly specialized entry routes with small numbers of places, so as to artificially inflate the points requirements and hence the prestige and attractiveness of the courses to high achievers.

Current reforms

There has been a view in recent years that the transition from second to third level education is not working well and that this has a negative effect on both the quality of students' learning experience in senior cycle and their preparedness for Higher Education. Announcing the first steps towards implementing a number of reforms intended to address this, the Minister articulated the problem thus: 'The Leaving Certificate has been captured by the points system. And the points system has distorted behaviour at second level' (Quinn, 2013). The reforms, referred to as the Transitions agenda and operating under the by-line 'Supporting a Better Transition from Second-Level to Higher Education', identified three key directions for action: first, addressing any 'problematic predictability' that might be identified in an independent external review of the Leaving Certificate examinations; second, changing the Leaving Certificate grading system by reducing the number of distinct grades available at each level from 14 to 8; third (and what the Minister referred to as the 'real problem'), reversing the explosion in the number of increasingly specialized entry routes for courses at third level, which was seen as having artificially increased competition in the points market.

The assessment process

Nature of assessments

Some subjects have a terminal written paper only, but the majority have more than one assessment mode. Languages other than English have a listening comprehension test, an oral examination and a written paper that tests both reading comprehension and written production. Geography involves a report on fieldwork activities, history a research project, technological subjects involve both a practical skills test and a coursework project in addition to the written paper, art has four components, three of which are practical and so on. Business and science subjects currently have a written paper only, but there are plans for additional components as syllabi change. For example, the SEC, with the assistance of the National

Council for Curriculum and Assessment, is currently carrying out a trial of arrangements for practical assessment in the sciences.

All components are externally assessed; teachers play no part in the assessment of their own students for certification purposes, other than to supervise and authenticate coursework. There are no plans to change this position.

Written papers are usually between two and three hours' duration. Multiple choice items are rare, and the examinations largely consist of short-answer items and extended-response items of varying lengths. Examinations in all subjects may be taken through English or Irish (other than the examinations in the subjects English and Irish and those in the 'non-curricular languages').

A selection of questions from examination papers is included in Annex 1.

Preparation of examination papers

The preparation of examination papers is the responsibility of the Chief Examiner for each subject, a member of the permanent staff of the SEC. These chief examiners are subject experts and assessment specialists. They may also be responsible for examinations in subjects outside their own areas of specialist expertise, as subjects with small candidatures cannot justify having a full-time specialist. In these cases, a subject specialist is appointed on part-time contract as a Deputy Chief Examiner.

The Commission appoints drafters and setters of examination papers, including coursework briefs and practical examinations. These are usually experienced teachers and examiners who carry out this work on contract. They also prepare draft marking schemes and assessment grids, which identify the content area and intended cognitive objective tested by each item, so as to promote alignment with the intended weightings of these objectives. Checks for accessibility, potential bias and so on are carried out. The papers are reviewed by nominees of the universities, who may make recommendations to the chief examiner. At a late stage, a person who has had no involvement in the preparation of the paper works through it in the same manner as a candidate would and makes observations. Full details of the process for developing examination papers are given in the Commission's *Manual for Drafters, Setters and Assistant Setters* (SEC, 2009).

Examinations are not pre-tested, and all candidates take the examination on the same day. If there is any suspected leak of the content of a paper, it is withdrawn and replaced with a contingency paper, which

will have been independently prepared by the same means. All examination papers are published on the Commission's website on the day they are taken, where they are freely available to the public. The content of the examination papers is the subject of considerable comment in the media – seemingly much more so than in most other countries (Baird *et al*., 2014).

School-based assessment (coursework) – task specification

All coursework is externally set and marked by the SEC. The task briefs for such coursework are prepared in a similar way and subject to similar quality assurance as the written papers.

Marking students' work

Marking completed examination papers

Written papers are marked over a period of 26 days by examiners appointed by the SEC. They are almost always qualified teachers of the subject involved. On-screen marking is being gradually introduced.

Examiners are trained at a conference that is usually of two days' duration. In addition to having the marking scheme explained to them, they mark exemplars and discuss these with each other and their advising examiners (team leaders). These advising examiners, along with the chief advising examiner, form the chief examiner's advisory team and play a key role in quality assurance. They will have met at a pre-conference to assist the chief examiner in finalizing the marking scheme and ensure that they are all in agreement. They meet again at a post-conference – a meeting that occurs shortly after marking gets underway and that serves a critical function in the standard setting process.

At least 5 per cent of the work of all examiners is monitored by their advising examiners. Monitoring involves completely re-marking the script concerned and giving advice and instructions to the examiner based on this. The examiner has no control over which scripts are selected for monitoring.

Although not their main purpose, the viewing and appeals processes that occur after provisional results are issued serve as an additional layer of quality assurance. If these reveal possible problems with the work of an examiner, this is investigated and, if necessary, the examiner's entire batch is re-marked.

Marking school-based assessment (coursework etc.)

Coursework examiners are selected and trained in a manner similar to written examiners. Depending on the subject, coursework may be sent to the SEC for marking or may be marked by examiners visiting the school.

In either case, the work of these examiners is monitored in a similar way to that of written examiners. Oral examinations in language subjects are carried out by external examiners who visit the school to conduct one-to-one interviews with the candidates and mark them. These examinations are audio-recorded to allow monitoring and, if necessary, moderation of the work of the examiner, and to be available in the event of an appeal.

Standard setting process

Conceptualization of 'standards'

The term 'standards' is often used in different ways and, for the purposes of this case study, the taxonomy of phenomenal, causal or predictive, as proposed by Newton (2010), has been adopted. There is no official document in Ireland that explicitly states what the Department of Education and Skills or its agencies mean by the term 'examination standards'. Nevertheless, it is clear from the discourse surrounding examination standards that the commonly held view reflects a phenomenal conceptualization – certainly in the context of the maintenance of standards from year to year in a given subject. The assumption is that if two students obtain the same grade in a given subject in two different years, this ought to mean that they have displayed the same level of subject competence. While a causal definition seems quite alien to the discourse, the predictive is a little less so. However, any such predictive interpretation seems to be of a consequential rather than definitional nature. That is, while an assertion that standards are falling may be elaborated on by an assertion that the students with a particular grade are not doing as well in higher education as they used to (i.e. standards are compromised in a predictive sense), it seems likely that this is being considered to be a consequence of those students not being as well prepared as before, which is essentially a phenomenal articulation. Certainly, those who are most critical of what they see as grade inflation have an implied definition of standards that is phenomenal. For example, the website stopgradeinflation.ie defines the problem thus: 'Grade inflation is a trend over time of better grades being awarded in educational qualifications that is not matched by real improvements in learning' (Network for Irish Educational Standards, 2007). This definition suggests that the grades should consistently reflect the degree to which the learning has been successful – presumably by reference to some idealized permanent and objective yardstick of successful learning.

A number of official documents indicate that the SEC is responsible for maintaining standards within a given subject over time. For example,

a document on standard setting on the Commission's website states: 'Once these performance standards have been tested, reviewed and fully established, we then seek to ensure that the standards remain consistent over time' (SEC, 2016).

Moving on to the standard setting approach and the techniques used to achieve it, these clearly fall into the category that Newton referred to as attainment-referencing, using a combination of expert judgement and statistical information. Indeed, the Commission itself states this in a manner that follows Newton's description closely (SEC, 2016).

Determining grades

Standard setting is conducted on a subject-by-subject basis and separately at each level within a subject. There is no aggregation of subjects in the certification of the award, although HEIs aggregate the grades for tertiary entry purposes. Grades awarded in the examination correspond to a predetermined percentage range of the marks available. That is, the grade boundaries are fixed. Furthermore, there is no provision for applying any kind of scaling transformation to the raw scores. The raw mark therefore determines the grade in a pre-ordained fashion that is fixed over time and across subjects. This poses considerable challenges for maintaining consistency in grading standards over time, since it is impossible to guarantee (without pre-testing items) that a particular year's set of examination questions will be identical in demand to the set used in any other year. This grading dilemma is resolved by embedding a standard setting process within the marking process itself. That is, if there are indications that the marking process is producing a grade distribution that is inappropriate in the context of statistics from previous years and the levels of achievement being observed, adjustments to the marking schemes are used to achieve changes in the distribution of the raw marks and hence the grades. In essence, the procedure is as follows:

- the marking scheme prepared in advance of the examination is a draft and is expected to remain fluid until the standardizing process is complete
- preliminary adjustments may be made to this draft after the examination is taken and before the examiners receive their training
- after training, examiners mark a sample of scripts. Data from this process is analysed, consideration is given to any unforeseen issues that may have arisen, and qualitative assessments of the standard of work encountered are made

- if necessary, the draft-marking scheme is adjusted, so as to ensure that the combined effect of the examination paper and marking scheme represents a comparable standard to that of previous years. All scripts are then marked in accordance with the revised scheme.

More detail on the procedure is given in SEC (2016).

In deciding what, if any, adjustment should be made to the marking schemes, the linking process is less formal than at the script scrutiny meetings that are used in England. Historical reference scripts are generally not used for comparison, and judgements of comparability are therefore more implicit, as the senior team is not making judgements based on direct script comparisons, but instead based on their accumulated knowledge and experience of examination standards. Changes to the size of the cohort or to the proportion taking the examination at each level are considered when evaluating the quantitative information coming from the emerging grade distribution, although actual data on prior achievement of the cohort concerned are not available. In reality, the 'similar cohort adage' (Newton, 2011) is a dominant influence – examiners seem to accept that the judgemental task involved in aligning boundary standards across different examinations cannot be achieved with the level of precision required, and, in the case of large cohorts at least, do not challenge the logic that the statistics should be the dominant influence in the absence of an identifiable systemic change. One could reasonably say that, in the case of large subject cohorts, expert judgement is being used as a check rather than the main influence: if the statistics suggest a course of action that is reasonable in terms of the quality of work they are observing, the subject experts will take that action. In the case of smaller subject cohorts, expert judgement becomes the more dominant influence on any decisions to adjust the marking scheme, as the statistical information is less reliable.

The chief examiner for the examination is ultimately responsible for the final decisions on the marking scheme. Nevertheless, Commission staff monitor the statistics closely, too. If an emerging distribution is too far out of line with those of recent previous years, the chief examiner will not be allowed to proceed without producing a convincing explanation – supported by evidence – as to why it should be allowed to stand.

Public debates related to the Leaving Certificate examination and standards

Concerns about examination standards can generally be considered to be related to some form of comparability of standards. As this is a

state examination with only one provider, comparability across different boards does not arise. Three other forms of comparability of standards are relevant: comparability over time; comparability across subjects; comparability across levels (Higher versus Ordinary). All of these have received some degree of attention in the public discourse. Other forms of comparability that impinge on policymakers and end-users – such as comparability of qualifications across countries – rarely receive much public attention.

The most notable issues related to examining standards that have been raised in recent years are as follows:

- a concern that the examination is not testing the right kinds of skills (insufficient emphasis on higher order thinking)
- grade comparability over time (a concern that examination/educational standards are falling)
- grade comparability across subjects, especially in the context of a tertiary entry system that effectively treats grades received in different subjects as equivalent;
- comparability of standards across levels (Higher versus Ordinary)
- the continued fitness for purpose of the current standard setting methodology.

Testing the right skills

The most dominant issue of concern in recent public discourse is only indirectly a matter of standards. It is the extent to which the examination is testing the right kinds of skills. As in other countries, it is frequently asserted that the examinations place too much emphasis on knowledge recall and not enough on higher order thinking skills. While there is by no means agreement as to the extent of this problem, there is a general consensus that the examinations would benefit from an increase in such emphasis. It may be noted that while an acceptance that higher order thinking is underemphasized is generally associated with a belief that students are not adequately prepared for Higher Education or certain forms of employment, it does not necessarily imply a belief that standards have fallen. Rather, it is more associated with the view that the kinds of knowledge and skills that may have been adequate in the past are no longer adequate. The recent review of problematic predictability carried out as part of the Transitions agenda included these two recommendations:

v. Consideration should be given to placing more emphasis upon the assessment of higher order thinking skills in the examinations, in keeping with international trends in assessment.

vi. A more regular programme of revision of syllabuses is needed for the Leaving Certificate examinations to remain current. This is important for keeping up with improvements in assessment design (such as assessing more higher order thinking skills), as well as syllabus content (Baird *et al.*, 2014).

There is a commitment from the relevant agencies to address these recommendations, but the challenges are considerable. They include:

- how to retain high reliability in marking items that genuinely test higher order thinking
- lack of clarity about what 'higher order thinking' means in the context of particular subjects (e.g. what does such an emphasis look like in an L3 language examination?)
- the degree of formal notification and lead time required for a significant change in examination emphasis or for new syllabi
- the challenge for teachers and students, and the consequent significant impact such changes might have on grade distributions (especially in the context of fixed grade boundaries).

Grade comparability over time

Claims of grade inflation are as common in Ireland as elsewhere (e.g. O'Grady, 2009; *Irish Times*, 2004). Nevertheless, counterbalancing views – arguing that grade improvements are due to factors such as more focused teacher and student engagement in a highly competitive higher education entry market – sometimes also receive an airing in the media (e.g. Healy, 2015). Faulkner *et al.* (2010) used a stable reference test in mathematics to examine the mathematical competence of incoming students to the University of Limerick over a ten-year period. While this showed that the average mathematical competency of the entry cohort had declined, this was accounted for by the changing Leaving Certificate grade profile of entrants, and the performance on the reference test of students with the same Leaving Certificate mathematics grade did not show a statistically significant change over the period. On the other hand, O'Grady (2009) had noted that mathematics was the subject that appeared to have been least

affected by grade inflation in the period he examined, so the conclusions of Faulkner *et al*. might not generalize to other subjects.

The difficulties associated with identifying and measuring changes in achievement over long periods of time are well rehearsed in the literature (e.g. Newton *et al*., 2007). Nonetheless, this problem is arguably less important than fluctuations or drifts over short periods of time. Students rarely use examination outcomes to compete with those who have taken examinations decades previously; they are largely competing with other candidates of the same year or only a few years apart. Accordingly, it is reasonable for authorities to focus on ensuring stability in standards over short time periods, which is a more tractable problem. Though their examinations receive annual criticism on a number of fronts, the means used by the SEC to maintain consistency in standards from year to year has not of itself been subject to much public criticism. Nonetheless, a number of more specialized sources have identified the need for marking schemes to serve this comparatively unusual purpose as a potential barrier to improving the quality of the examinations (Baird *et al*., 2014; Newton, 2014; SEC, 2012).

Grade comparability across subjects

Comparability of grades across subjects has long been the focus of discussion in Ireland (e.g. Commission on the Points System, 1999; Kellaghan and Millar, 2003; Hyland, 2011). For example, the potentially detrimental effect on the uptake of the physical sciences, which are perceived to be relatively difficult to score well in, have been a cause of concern (e.g. Task Force on the Physical Sciences, 2002). While there is some evidence of differences in grading standards across subjects, it is not clear that such discrepancies have a substantial effect on subject choice. For example, Millar (2014) found no evidence within examination data sets of strategic subject choices by candidates who might be expected to be highly motivated. On surveying students' reported reasons for their subject choices, Smyth and Calvert (2011) found that, while a belief that a subject might be easy to do well in was an influence, it ranked behind interest in the subject and a belief in its value or necessity for a future career. Guinan (2001) had similar results.

Nevertheless, comparability of grading across subjects is considered a matter of fairness, especially when grades are to be aggregated for the purposes of making tertiary entry decisions. Kellaghan and Millar (2003) analysed grading practices in the Leaving Certificate examinations of 1996,

2000 and 2001. They found, unsurprisingly, that grade distributions differed significantly across subjects. However, considering these distributions in the context of the prior academic achievement profile of their cohorts and their current academic achievement profile (based on a subject-pairs analysis), they noted that subjects with academically stronger cohorts tended to have better grade distributions, though not better by as much as one might expect. They summarized it thus:

> An alternative explanation [of the findings] is more complex, and proposes that examiners reach a kind of compromise in grading, in which they attempt to balance examinees' overall academic achievement, the nature and demands of the syllabus they have followed, and the need to provide an acceptable distribution of grades for every subject, at both Higher and Ordinary level. The effect of the compromise reflected in the grades awarded in the Leaving Certificate examination is that the grades of high achieving candidates are lower than one would expect on the basis of their overall achievement, and the grades of low achieving candidates higher (Kellaghan and Millar, 2003).

The analyses were repeated by Millar (2014) on the 2013 Leaving Certificate examination. The results were generally similar to those of the previous study. At the request of the Transitions steering group in 2015, Millar also carried out an IRT-based analysis (unpublished), following a similar methodology to Coe (2008). The primary purpose of this study was to evaluate the existing linkage implied by the points system between Higher and Ordinary level grades within subjects, but in doing so it also generated ability estimates corresponding to each grade in each subject. In terms of the rank ordering of subjects, the results were again similar. In general, the ordering of subjects in this respect is similar to the ordering found in such studies in other countries over a long period of time, as noted by Pollitt (1996) and others.

While a measured and nuanced literature on the topic exists, the public discourse surrounding this issue has been relatively unsophisticated. For example, it is periodically suggested that all subjects should have the same grade distribution imposed upon them, notwithstanding that this would demonstrably make existing discrepancies in grading severity (as measured by subject pairs analysis or IRT methods) worse rather than better.

Despite the known difficulties with assuming that the same grades in different subjects should be treated as equivalent, there is little appetite

for deviating from this assumption in any revision to the tertiary entry system. In the recent discussions on changes to the grading system and the calculation of points for tertiary entry, there was no appetite for a composite measure that would incorporate a scaling procedure to account for subject differences, such as the Australian Tertiary Admission Rank. scaling and other statistical techniques have not been a feature of the grading or aggregation processes to date, and the simplicity of the current system is valued. It seems likely that HEIs and other end users will continue to treat grades in different subjects as though they were equivalent, while remaining aware that they are not.

Linking standards across levels

Paradoxically, one standards-related issue of significant current interest to the SEC has received little public attention. Until 2016, the linkage between grades awarded at Higher level and those awarded at Ordinary level was a construct of the HEIs as end users of the certification. It had no standing in the eyes of state agencies. There was therefore no onus on the SEC to ensure that examining standards at the two levels reflected this linkage. However, in a very significant policy change in 2015, the Department directed that examining and grading standards should in future be aligned to the linkage implied by the new points system. This posed a significant challenge not faced heretofore. While the IRT-based research on existing data carried out on behalf of the Transitions group concluded that, on average across all subjects, grading standards are not too far distant from those implied by the old points scheme, it also identified large differences between subjects in this respect. If the new policy is to be made a reality, significant realignments of standards will be required in many subjects.

Given these implications, a paper was commissioned from Newton (2014) by the SEC to identify and explain the issues involved, and the implications for examination and specification design. This chapter briefly explored conceptualizations of standards, comparability, and linking in the context, identified possible strategies for trying to realize the stated objective, and identified advantages and disadvantages of each. While recognizing that there is no perfect solution, this chapter suggests that, in this context, conceptualizing comparability as an approximation to a linking relationship is reasonable, and that various techniques could be used to enhance such comparability. In particular, common item approaches would seem to have the most potential for making the linkage more robust, with other techniques either playing a subsidiary role or an alternative role in cases where no common items of components are present.

Based on Newton's paper (2014), the SEC is preparing proposals for consideration by the Department of Education and Skills as to how best to ensure that the stated alignment becomes a reality, backed up by as robust a linking procedure as is feasible. However, it is clear that none of the available procedures that have even a moderate degree of robustness could be implemented unless the current methodology for standard setting in the Leaving Certificate is changed.

Limitations of the standard setting methodology

As previously noted, the constraint that grade boundaries are fixed and raw scores are not subject to any standardizing transformations necessitates embedding the standard setting process within the marking process. While this approach has arguably served its purpose well enough to date, there are limits to what it can achieve. Apart from its inefficiency and the fact that it cannot achieve the same level of precision as could be achieved by scaling scores or adjusting grade boundaries, it has some negative consequences. In preliminary internal research on examination predictability, carried out in advance of the external study commissioned in the context of the Transitions reforms, chief examiners expressed the view that the need to use marking schemes to keep the grade distribution comparatively stable over time was hindering their capacity to be innovative in their questioning and was thereby contributing to predictability (SEC, 2012). The external review itself also identified the standard setting procedure as potentially hindering the capacity of the examination papers and marking schemes to seek and reward higher order thinking skills:

> Adjusting the marking distribution by altering the marking scheme is more manageable if the changes relate to factual issues. More subtle judgements regarding higher order skills would be more difficult to revise in a reliable manner at a time when the examination system is under a great deal of time pressure and the expectations for marking reliability are high. Without the constraint of fixed cut-scores, it may be more straightforward to achieve a better balance between the assessment of knowledge and higher order thinking skills (Baird *et al.*, 2014).

The report also noted that the standard setting procedure – necessitating as it does questions with high mark tariffs and marking schemes with built-in flexibility – rendered the marking schemes less transparent than they might otherwise be:

> Any changes to the marking schemes to make them more transparent could have implications for the manageability of fixed grade boundaries (cut scores) in the Irish Leaving Certificate. Thus, there are decisions to be taken about whether marking schemes can be changed in this way whilst maintaining the current standard-setting system (Baird *et al.*, 2014).

Another challenge is presented by the new requirement to link grades across levels. The most reasonable form of linking to seek in this context – in the longer term at least – is through the use of common components and common items. In practice, the requisite data from the marking of the common elements could not be processed and analysed in time to feed into the marking process for the remaining elements, making it impossible to use a standard setting process that is embedded in the marking process to realign one or both distributions.

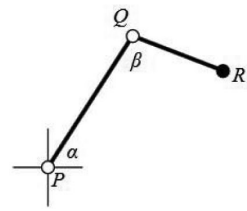
The lack of a tradition of scaling raw scores to suit a particular purpose has led to some other challenges. Some subjects in the Leaving Certificate have common-level components. For example, in the examinations of modern languages, the oral component is assessed at a common level and later combined with other components assessed at Higher and Ordinary levels. Not surprisingly, Ordinary-level candidates perform less well on the common-level oral test while Higher-level candidates score better. While it might seem natural to standardize the marks to the level of the candidate by applying a transformation to one or both sets of scores, this is not done. Instead, the chief examiners ameliorate the effect by compensating in the standards required in the other components. While this is generally effective, it is arguably not an ideal way to deal with what is essentially a numerical problem with a numerical solution.

The SEC is currently finalizing proposals to put to the Department of Education and Skills outlining its views as to the standard setting procedures that would be appropriate in the context of what the education system now seeks from the examinations as a result of the Transitions reforms. No decisions have yet been taken, but it is quite possible that standard setting in the state examinations in Ireland will look different in a few years' time from how it looks today.

Annex 1: Some Leaving Certificate examination items

Question 8 **(75 marks)**

The diagram is a representation of a robotic arm that can move in a vertical plane. The point P is fixed, and so are the lengths of the two segments of the arm. The controller can vary the angles α and β from 0° to 180°.



(a) Given that $|PQ| = 20$ cm and $|QR| = 12$ cm, determine the values of the angles α and β so as to locate R, the tip of the arm, at a point that is 24 cm to the right of P, and 7 cm higher than P. Give your answers correct to the nearest degree.

(b) In setting the arm to the position described in part **(a)**, which will cause the greater error in the location of R: an error of 1° in the value of α or an error of 1° in the value of β?

Justify your answer. You may assume that if a point moves along a circle through a small angle, then its distance from its starting point is equal to the length of the arc travelled.

(c) The answer to part **(b)** above depends on the particular position of the arm. That is, in certain positions, the location of R is more sensitive to small errors in α than to small errors in β, while in other positions, the reverse is true. Describe, with justification, the conditions under which each of these two situations arises.

(d) Illustrate the set of all possible locations of the point R on the coordinate diagram below. Take P as the origin and take each unit in the diagram to represent a centimetre in reality. Note that α and β can vary only from 0° to 180°.

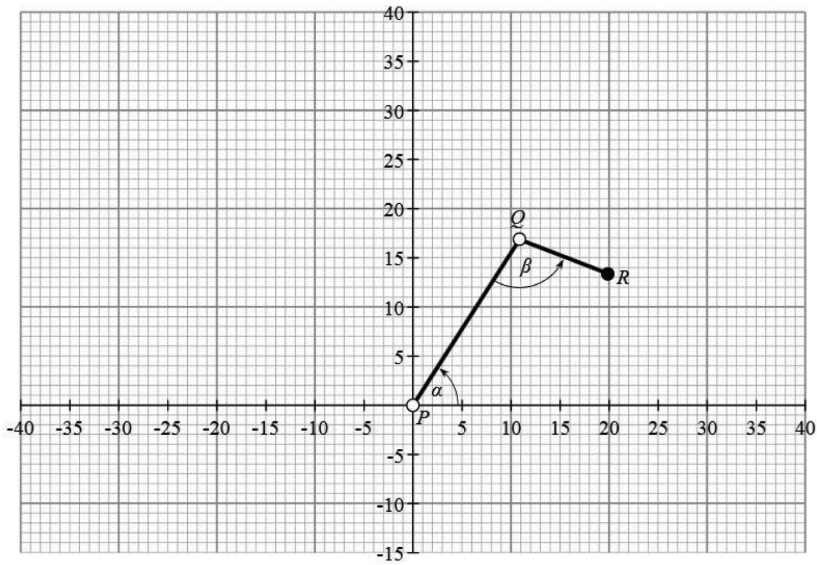


Figure 9.1: Higher Level Mathematics 2012 (Project maths – Phase 3)

Question 4

A. Map Skills

Draw an outline map of a **Continental / Sub-Continental** region (not in Europe) that you have studied.

On it, show and name each of the following:

- The outline of a named feature of relief (upland or lowland)
- A named urban centre
- The outline of a named agricultural **or** industrial region
- The course of a named river.

[20m]

B. European Union

Examine the impact of the expansion of the European Union on member states, with reference to both economic and social impacts.

[30m]

C. Secondary Economic Activity

Examine the development of secondary economic activity in an **Irish** region that you have studied, with reference to any **two** of the following factors:

- Raw materials
- Transport
- Labour
- Markets.

[30m]

Figure 9.2: Higher Level Geography 2016

QUESTION A

(i) Outline, in your own words, the insights Andrew Dickson shares about Shakespeare's play *The Comedy of Errors* in the written text above. (15)

(ii) From the four posters in the above text, choose the poster that you think is most effective in advertising a production of the play, *The Comedy of Errors*. Explain your choice with reference to the written text and the content and visual appeal of the poster. (15)

(iii) The writer suggests that Shakespeare's plays have timeless and global qualities. Do you agree with this view? Support your answer with reference to the above text (written and visual) and your own experience of at least one Shakespearean drama, other than *The Comedy of Errors*. You may refer to written texts, stage productions or films. (20)

Figure 9.3: Higher Level English 2016

SECTION II: PRODUCTION ÉCRITE (100)

Répondez à *trois* questions – la Question 1 et deux des Questions 2, 3 et 4.

N.B. LA QUESTION 1 EST OBLIGATOIRE.

Q.1. Répondez à *(a)* **ou** à *(b)*. (40 points)

(a) Selon Jonathan, « *son futur métier nécessitera une bonne connaissance des langues étrangères* ». [Section I, Q.1]

Dans les écoles en Irlande, on attache plus d'importance aux mathématiques et aux matières scientifiques qu'aux langues étrangères.

Êtes-vous d'accord ?

(90 mots environ)

OU

(b) Selon le texte de la Section I, Q.2, Mia « *accéda à sa boîte mail et parcourut tous les courriels…….* »

Vous ouvrez votre boîte mail un jour et vous trouvez un courriel intéressant ! Vous décidez de répondre. Racontez ce qui s'est passé ensuite. (*Votre récit peut être réel ou imaginaire.*)

(90 mots environ)

Figure 9.4: Higher Level French 2016

References

Baird, J., Hopfenbeck, T., Elwood, J., Caro, D. and Ahmed, A. (2014) *Predictability in the Irish Leaving Certificate*. Oxford: Oxford University Centre for Educational Assessment. Online. https://goo.gl/dSFzAy (accessed 10 June 2018).

Coe, R. (2008) 'Comparability of GCSE examinations in different subjects: An application of the Rasch model'. *Oxford Review of Education*, 34 (5), 609–36.

Commission on the Points System (1999) *Final Report and Recommendations*. Dublin: Stationery Office.

Department of Education and Science (1995) *Leaving Certificate French Syllabus: Ordinary and higher levels*. Dublin: Stationery Office.

Faulkner, F., Hannigan, A. and Gill, O. (2010) 'Trends in the mathematical competency of university entrants in Ireland by leaving certificate mathematics grade'. *Teaching Mathematics and its Applications*, 29 (2), 76–93.

Guinan, A.M. (2001) 'Who, What and Why ... Subject Choices for Senior Cycle in a Second Level School'. Master's Degree thesis. National University of Ireland Maynooth. Online. https://goo.gl/SPhs5G (accessed 10 June 2018).

Healy, C. (2015) 'Is the Leaving Cert getting easier?' *thejournal.ie*. 2 June. Online. http://jrnl.ie/2132187 (accessed 10 June 2018).

Hyland, Á. (2011) *Entry to Higher Education in Ireland in the 21st Century*. Ballsbridge, Dublin: Higher Education Authority. Online. https://goo.gl/gGPyMG (accessed 10 June 2018).

Irish Times (2004) 'Concern at falling standards in science and maths'. *Irish Times*. 17 August. Online. https://goo.gl/7Fmmku (accessed 10 June 2018).

Kellaghan, T. and Millar, D. (2003) *Grading in the Leaving Certificate Examination: A discussion paper*. Dublin: Educational Research Centre.

Millar, D. (2014) 'An exploration of variation in subject grading, student subject selection and outcomes in the leaving certificate examination'. Unpublished report.

Network for Irish Educational Standards (2007) *Grade Inflation: Summary of Papers 1–5*. Online. http://stopgradeinflation.ie/wp-content/uploads/2017/09/Grade-Inflation-A-Summary-papers-1-to-5.pdf (accessed 19 July 2018).

Newton, P.E. (2010) 'Contrasting conceptions of comparability'. *Research Papers in Education*, 25 (3), 285–92.

Newton, P.E. (2011) 'A level pass rates and the enduring myth of norm-referencing'. *Research Matters*, Special Issue 2, 20–6.

Newton, P.E. (2014) *Comparability of Standards between Higher and Ordinary Level Leaving Certificate Examinations*. London: Institute of Education.

Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) (2007) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority. Online. https://goo.gl/6kqoXe (accessed 7 June 2018).

OECD (Organisation for Economic Co-operation and Development) (2016) 'Educational attainment of 25–64 year-olds'. Online. stats.oecd.org (accessed 9 October 2016).

O'Grady, M. (2009) *Grade Inflation in the Leaving Certificate Examination 1992–2006*. Online. https://goo.gl/aTzPZ8 (accessed 11 June 2018).

Pollitt, A. (1996) 'The "difficulty" of A level subjects'. Unpublished report for the University of Cambridge Local Examinations Syndicate.

Quinn, R. (2013) *Key Direction for Change*. Video. Online. https://vimeo.com/ncca/review/68234791/0164dade3c (accessed 11 June 2018).

SEC (State Examinations Commission) (2009) *A Manual for Drafters, Setters and Assistant Setters*. Online. https://goo.gl/8j6kDg (accessed 11 June 2018).

SEC (State Examinations Commission) (2012) Unpublished draft report of the SEC working group on predictability in the Leaving Certificate examination.

SEC (State Examinations Commission) (2016) *Setting and Maintaining Standards in the State Examinations*. Online. https://goo.gl/gh4Wde (accessed 11 June 2018).

Smyth, E. and Calvert, E. (2011) *Choices and Challenges: The transition from junior cycle to senior cycle education*. Dublin: ESRI and Liffey Press.

Task Force on the Physical Sciences (2002) *Report and Recommendations*. Dublin: Stationery Office. Online. https://goo.gl/K1a4qE (accessed 11 June 2018).

Does the Leaving Certificate reward LOT (lower order thinking) rather than HOT (higher order thinking)?

Áine Hyland

This is a very welcome and informative chapter on setting standards in the public examinations system in Ireland. It describes the approach taken by the State Examinations Commission (SEC) in setting and marking public examinations and recognizes both the strengths and weaknesses of the system.

The chapter rightly points out that the public examinations system is held in high esteem by the general public in Ireland. The Leaving Certificate (LC) is used by all Irish higher education institutions to select incoming students; given that Ireland has one of the highest transfer rates of students from second to third level in the OECD, this is a very important vote of confidence in the system.

The chapter reflects the open and transparent approach taken by the SEC in recent years, which has been widely welcomed. Candidates can now view their marked examinations scripts following receipt of their LC results. Students also have access (post hoc) to the marking schemes used by the examiners. This has resulted in the unintended consequence of an undue focus by students and their teachers on marking schemes and their application. In the final year of secondary education, focus is often more on examination techniques than on scholarly engagement with the subject.

The chapter addresses a number of the common criticisms of the Irish public examinations system. These include the inconsistency of standards across different subjects; the issue of grade inflation; and an inadequate emphasis in the LC on rewarding higher order thinking.

The fact that 'standard setting is conducted on a subject-by-subject basis' indicates one of the weaknesses of the system. This can result in an inconsistency of approach by different chief examiners and can exacerbate what may already be an intrinsic difference in the academic levels of students taking different subjects.

The challenge of maintaining consistency in grading standards over time is also addressed. In order to ensure that there is no significant discrepancy between the distribution of grades from one year to the next, adjustments are made to the marking schemes to achieve changes in the distribution of raw marks and grades. Historical reference scripts are generally not used for comparison, and judgements of comparability are implicit and subjective. This prevents direct year-to-year comparison of standards and makes it impossible for an outsider to identify whether or not grade inflation has occurred. Having said that, there is no disputing the evidence that there is a high and consistent correlation between the results of students in the LC and their subsequent results in university examinations. This is a phenomenon that the critics of the LC marking schemes have failed to explain satisfactorily.

The chapter recognizes that 'the most dominant issue of concern in recent public discourse … is the extent to which the examination is testing the right kinds of skills' and acknowledges that 'it is frequently asserted that the examinations place too much emphasis on knowledge recall and not enough on higher order thinking skills'. The report by Baird *et al.* (2014) on 'Predictability in the Irish Leaving Certificate' recommended that consideration should be given to placing more emphasis on higher order thinking skills in the examinations, in keeping with international trends in assessment. In the view of this commentator, this is the most pressing issue that needs to be addressed by the SEC in any reform of the system. While recognizing the challenges involved in assessing higher order thinking, the stakes are high. Future populations need to have critical, analytical, problem-solving and creative skills to enable them to engage with and resolve the many challenges facing society – whether political, social, cultural, economic or work-related. An examination system that rewards knowledge recall to the detriment of these higher order skills will no longer serve society adequately, if it ever did.

Reference

Baird, J., Hopfenbeck, T., Elwood, J., Caro, D. and Ahmed, A. (2014) *Predictability in the Irish Leaving Certificate*. Oxford: Oxford University Centre for Educational Assessment. Online. https://goo.gl/dSFzAy (accessed 10 June 2018).

The delicate task of standard setting

Michael O'Leary

In this chapter Hugh McManus provides a succinct description of the essential elements of the Leaving Certificate (LC) examination system in Ireland. Reading it serves to remind me that it is to the State Examinations Commission's (SEC) great credit that the papers for this high stakes examination are set, administered and graded with the minimum of fuss despite the intense media focus during the June examination period each year and when the results are published every August. This may be one of the reasons why public support for the LC has been high in the past and many still contend that the procedures put in place to guarantee the anonymity of those taking and marking a set of standardized exams constitutes a level of fairness that is difficult to replicate with most other forms of assessment. With that in mind, the delicate task of maintaining a balance between the public's confidence in the LC and informing the public about the implications of measurement error for setting and maintaining standards is worth considering. McManus refers to the challenge of maintaining high reliability in marking but does not elaborate on how reliability is currently assessed. The basis for judgements during the standard setting process that an emerging distribution of grades is 'significantly' out of line with those from previous years of the exam is also unclear. We know that in any given year, while approximately 18 per cent of LC papers sent to be rechecked are upgraded, this constitutes just 0.44 per cent of all LC grades (the equivalent figures for total GCSE, AS and A levels are very similar). Perhaps this is evidence for relatively high levels of consistency in the marking process, but in the absence of any other data it is difficult to be sure. Studies to calculate inter-rater reliability coefficients and standard errors of measurement for LC subjects should be undertaken and published. Akin to how sampling error statistics are used when communicating about the outcomes of opinion polls, measurement error statistics could be used to make LC grades seem less definitive as a measure of achievement than is currently the case. Other commentaries in the public arena contend that the LC is old fashioned (e.g. Baird *et al*., 2015) and needs to be reformed in tandem with changes in curricula that better reflect the knowledge, skills and dispositions required for the world of further education and work in the twenty-first

century. With these in mind, it is good to see McManus reiterating the very important point that the standard setting process used for the LC may be acting as a barrier to the incorporation of exam questions that tap higher order thinking skills.

The conversion of LC grades to a points system (aka Central Application Office [CAO] points) for use in selecting students for entry to third level education in Ireland serves to highlight many problematic issues with respect to the standard setting process as McManus expertly elucidates in this chapter. He is correct in stating that grades across different subjects, levels of subjects and different years of the LC are treated as if they were equivalent by higher education institutions even though in reality they may not be. With that in mind, a study of why just over 11 per cent of those taking chemistry achieved the highest LC grade in 2017 compared to just under 5 per cent of those taking biology would be illuminating. An investigation into the grade distributions for higher level mathematics from 2012 onwards would also be of interest given the sharp increase in the numbers taking up the option incentivized by an additional 25 bonus CAO points. For example, there were fewer B grades than normal in 2012, while there was also an increase in the proportion of C grades achieved. In 2013 and subsequent years, there was a sharp increase in the proportion achieving Grade D and an evening out at Grades B and C.

It is good to read in the chapter that the SEC is reviewing approaches to common item linking as a means of addressing some of these problems. However, it is also sobering to read that there is little appetite among some key stakeholders for the use of more sophisticated scaling procedures. McManus is not alone in believing that while the current LC standard setting procedures have served the Irish system reasonably well in the past, planned reforms of the LC means that a more robust and transparent system needs to be put in place as a matter of urgency.

Acknowledgements

I am grateful to Dr Gerry Shiel, Educational Research Centre, Dublin, for his feedback on an early draft of this commentary and for his observations about LC mathematics' grade distributions.

Reference

Baird, J., Hopfenbeck, T.N. and Caro, D. (2015) *Predictability in High-Stakes Assessment: Students' approach to learning*. Video podcast. Online. https://goo.gl/17SG38 (accessed 11 June 2018).

Chapter 10

Standard Setting in Queensland: The Queensland Certificate of Education

Matthew Campbell

Introduction

Queensland's system of senior assessment and tertiary entrance is currently in transition. While this chapter describes the existing system in broad terms, its main focus is on the new arrangements, which are currently in development. Readers should therefore be aware that the new system remains a work-in-progress at the time of publication.

Queensland is the second largest state in the federation of Australia. Its population is the most dispersed, with nearly half of approximately 4.8 million people residing in the south-east corner around the capital city of Brisbane, but with large population centres distributed across the entire state. For example, Cairns in the northern part of the state has a population of approximately 160,000, and is located approximately 1,700 kilometres from Brisbane. Each Australian state has constitutional responsibility for the delivery of education, although the Australian Government also influences education and schooling through the provision of funding to the state governments and directly to non-state schools and systems.

School education in Queensland is delivered across three main sectors: government schools, Catholic schools and independent schools. There are approximately 1,725 schools in the state, of which 278 are dedicated secondary (Years 7–12) schools, with a further 272 being combined primary and secondary. Students generally commence compulsory school education around the age of 5 entering into Prep, and must attend school until the age of 16 or the completion of Year 10 (whichever is earlier). Approximately 80 per cent of all students who commence secondary schooling in Year 7/8 progress to completion of Year 12 (the final year of secondary schooling), with approximately 29,000 Year 12 students (or 61 per cent of the cohort) applying to attend university post-school (Department of Education

and Training [Cwth], 2015). Government schools are administered by the Department of Education, under the leadership of the Minister for Education, with the bulk of funding for these schools allocated by the state government. This is by far the largest sector comprising nearly 1,250 schools across the state (Department of Education and Training [Qld], 2016a).

Catholic schools are independent of government, constituted under their own system of governance, with approximately 300 schools, comprising 18.33 per cent of the total school population and 60 per cent of the non-government school population (QCEC, 2016). Independent schools are individual schools constituted under their own board, with approximately 14 per cent of all students enrolled at these schools (QGSO, 2011). These schools may or may not be religious and receive the majority of funding from the Commonwealth government. In secondary schooling, approximately 39 per cent of students are enrolled in non-state schools (Department of Education and Training [Qld], 2016a).

The Queensland Curriculum and Assessment Authority (QCAA) is a statutory authority with responsibility for the development and revision of syllabi across Prep (preparatory year) to Year 12, support for the implementation of the Australian Curriculum in Queensland, and management of associated testing and assessment processes (2014 Education [QCAA] Act). The QCAA certifies student achievement of the completion of Year 12 with the Queensland Certificate of Education (QCE). The Queensland Tertiary Admissions Centre (QTAC), a public company established by a consortium of tertiary institutions, is responsible for the management of tertiary entrance processes, and from 2020 will calculate the tertiary admission rank for students in Queensland.

Understanding standards in the Queensland context

Queensland's current assessment approach for senior students (i.e. students completing high school and seeking tertiary entrance) is described as a system of externally moderated school-based assessment. Assessment is designed and executed by schools and teachers based on guidance contained in syllabus documents. Teacher judgements, based on standards presented in the syllabi, are reviewed through an external moderation and verification process. All assessment in the current system is standards-based. Teachers make judgements about the quality of student achievement with reference to predefined standards that describe how well students have achieved the objectives in syllabi. Predefined standards ensure that:

- students and teachers know what is expected for each level of achievement and can work together to achieve the best result for the student
- comparability from school to school can be achieved
- teachers can discuss standards with parents or carers when reporting a student's achievements.

Within the syllabus for each subject, objectives are grouped by dimensions and presented in a standards matrix which describes the standards for each dimension, expressed on an A–E grade scale. Teachers use the standards matrix first at the level of the individual assessment instrument; that is, through considering how students are progressing towards or already demonstrating achievement of final standards, and second, for decisions about overall achievement across a range of assessment instruments towards the end of the course. These decisions are on balance judgements about how the qualities of the student's work match the standards descriptors overall in each dimension. On completion of a senior secondary course of study, teachers award one of five levels of achievement.

The QCAA administers a system of social moderation designed to ensure that results recorded match the requirements of the syllabus. The aim of moderation is to ensure comparability – that is, students who take the same subject in different schools, and who attain the same standard through assessment programmes on a common syllabus, will be awarded the same level of achievement. This does not imply that two students who receive the same level of achievement have had the same collection of experiences or have achieved equally in any one aspect of the course. Rather, it means that they have, on balance, reached the same broad standard.

In the current and future systems, standards are used in three distinct but interrelated ways: standards of assessment, standards of learning and standards descriptor. A standard of assessment is defined as 'a fixed reference point used to describe how well students have achieved the outcomes or objectives in syllabi. The descriptions of standards of assessment, also referred to as reporting standards, are derived by groups of teachers and subject experts describing the actual differences in examples of student work'. They are statements that succinctly describe typical performance at each of the five levels (A–E), reflecting the cognitive taxonomy and objectives of the course of study. The standard of learning is understood as a 'statement of what students are expected to know and do by the end of key junctures of schooling (outcomes or objectives) and the scope of that learning (core content or subject matter)'. Finally, a standards descriptor

is 'a statement (or list of statements) that succinctly conveys the required quality of, or features in, student work in order to be awarded a particular standard of achievement'. These are defined within the marking guides and performance level descriptors within syllabi (QCAA, 2014:12).

University entrance examinations

In all Australian states and territories, senior secondary students seeking entrance to university are awarded a rank based on their achievements in their school subjects. In most jurisdictions, final subject results are derived from a combination of external and school-based assessment, with the external assessment results commonly used to scale the school-based assessment results (Blyth, 2014). Currently, in Queensland (and the Australian Capital Territory), students' final subject results are derived entirely from their achievements in school-based assessments. Assessment instruments devised by teachers, and the judgements they make about how well the students have learnt, are the major component of students' final results. In the new Queensland system, final subject results will be determined through combining student achievement on school-based assessment and one external assessment without the scaling of any assessment by another.

In Queensland, most students work towards a Queensland Certificate of Education (QCE), which is typically awarded at the end of Year 12 (although students may continue their studies post-school and be awarded a QCE when eligible). Currently, results in a student's subjects are used in the calculation of a tertiary entrance rank, known as an Overall Position (OP). Although entry to higher education can be achieved through multiple pathways, the most common direct entry for senior students in Queensland is via an OP rank calculated from their best five individual subject results and their school group performance on a common scaling test, the Queensland Core Skills (QCS) Test. In the revised assessment system, the current OP rank is to be replaced by the Australian Tertiary Admission Rank (ATAR), and students will no longer undertake the QCS Test. Instead, comparability in most subjects will be achieved through a combination of external assessment, and new processes requiring the endorsement of school-based assessment instruments, and confirmation of teacher judgements, generating a final subject result. Calculation of an ATAR will be based on the combination of subject results, with eligible students required to complete a minimum of four general (i.e. university preparation) subjects.

Queensland's current system of senior assessment was implemented in the early 1970s following widespread concern that the Senior Public Examination had become too focused on the tertiary entrance priority of determining academic excellence and was no longer fit-for-purpose in responding to the changing goals of senior secondary education resulting from increasing student retention to Year 12 (Radford, 1970; Clarke, 1987). A disconnect had emerged between the goals of senior secondary curriculum and the use of assessment for certification and tertiary entrance purposes. A number of subsequent reviews (the Scott Review of School-based Assessment [1978], Pitman [1987] and Viviani [1990] reports) modified and updated policies and practices concerning the use of assessment and standards in senior secondary education, but the reliance on school-based assessment has remained unchanged (Kelly, 2014). Therefore, since the 1970s the Queensland curriculum and senior secondary authorities (i.e. the QCAA and its predecessors) have not managed direct university entrance examinations.

Following a 2013 parliamentary inquiry into assessment methods used in mathematics, physics and chemistry (Queensland Parliament, 2013a), the Queensland Government appointed the Australian Council for Educational Research (ACER) to conduct a full-scale review of senior assessment and tertiary entrance processes (Queensland Parliament, 2013b). In the report released in 2014, it was found that while existing arrangements had served Queensland students well and remained fair and reliable, they would not be sustainable over the longer term (Matters and Masters, 2014). ACER recommended changes to achieve greater rigour and simplicity. These included:

- reducing the number of summative assessments to be undertaken by students in Year 12 (currently the assessment load for students in their final year of schooling can be as high as 40 assessments, in addition to the QCS Test)
- introducing subject-based external assessment in Year 12 that complements school-based assessments and is not used for scaling purposes
- increasing the rigour of the processes for external scrutiny of school-based assessment instruments and teacher judgements about student achievement
- separating the responsibilities for certifying a subject result from those associated with tertiary entrance (the QCAA currently performs both tasks) and replacing the OP with an ATAR.

In response, the Queensland government decided to introduce a revised system starting with students entering Year 11 in 2019. The new model will have the following features:

- students will complete three school-based assessments and one external assessment in most senior subjects, with the majority of students undertaking the equivalent of six subjects in their two years of senior schooling
- school-based assessment will contribute 75 per cent to a student's final subject result in most subjects, 50 per cent in mathematics and science subjects
- subject-based external assessments will be introduced in most subjects, but they will not be used to scale students' school-based assessment results in the derivation of a final subject result
- school-based assessment instruments will be endorsed by the QCAA before they can be used for summative purposes in schools
- QCAA will confirm the grades awarded by schools by reviewing a selected sample of student work for every subject in every school
- there will be no dedicated scaling test or examination used as a selection mechanism for post-schooling pathways. An ATAR will be derived from achievement across a broad range of learning achievements using a process of inter-subject scaling. It will be calculated from an eligible student's best five subject results with no compulsory inclusion of specific subjects. However, to be eligible for an ATAR a student must satisfactorily complete an English subject. One of the five subjects may be an applied learning subject that does not include an external assessment, or a competency-based vocational education and training certificate at a specified level.

The QCAA is in the process of implementing the government's policy by redeveloping its suite of syllabi, developing new processes to strengthen the quality assurance of school-based assessment, and implementing processes to support the introduction of subject-based external assessment.

The assessment process

This section focuses on the revised system of senior assessment set to commence in Queensland in 2019. In the new system, subjects undertaken in the senior curriculum will be renamed general or applied subjects. General subjects, currently known as Authority subjects, will cover subjects designed as preparation for university and higher education studies. Applied subjects will be aimed at preparing students for employment, or

technical education, and focus on applied learning and practices. All general subjects will be organized into four units. Units 1 and 2 will generally be foundational learning, allowing students to begin engaging with the course subject matter, and to experience the objectives of the syllabus. Units 3 and 4 will consolidate student learning, with the assessment results for these units contributing to the final subject result and tertiary entrance rank. Final results from a combination of five general subjects, or four general subjects and one applied subject or vocational qualification, will be used in the calculation of an ATAR.

Nature of assessments

Achievement in the QCE will continue to be reported using an A to E scale of achievement, with an accompanying numerical subject result used for tertiary entrance purposes. Overall achievement standards in subjects will be derived from a combination of school-based and external assessment, using a variety of complementary yet separate approaches to assessment. Approaches to assessment across the senior syllabi broadly reflect six assessment techniques as described below:

- project: a response to a single task, situation or scenario in a unit of work that provides for authentic opportunities for students to demonstrate their learning, comprised of at least two assessable components demonstrated in different contexts, to different audiences and through different modes
- investigation: the investigative process of locating and using information beyond a student's own knowledge, usually engaging with research and inquiry approaches to learning
- extended response: the interpretation, analysis, examination and/or evaluation of ideas and information usually in response to a provided stimulus, and may involve additional research
- performance: physical demonstrations of outcomes across a range of cognitive, technical, physical and/or creative and expressive skills, through the application of identified skills to either solving a problem, providing a solution or conveying meaning and intent
- product: the production of physical and virtual objects and representations through the application of cognitive, technical, physical and/or creative and expressive skills
- examination: the application of a range of cognition to provided questions, scenarios and/or problems, undertaken individually, under supervised conditions and in a set timeframe.

Examinations and external assessments

Subject-based external assessment is being reintroduced into the Queensland senior curriculum after more than 40 years. Within the new system, the term 'external assessment' refers to an assessment task undertaken by students at the end of a course of study but not assessing the full course of study; however, it is emerging that nearly all external assessment will take the form of an examination. Unlike other Australian jurisdictions (e.g. the Higher School Certificate in New South Wales and the Victorian Certificate of Education), Queensland's external assessments are generally not intended to assess content and skills across the entire subject but instead are focused on particular units or aspects of study. Students will complete four summative assessments across Units 3 and 4, with most syllabi providing for the external assessment to occur towards the conclusion of the school year, focusing on the last unit of work, or one of its topics. The nature and form of each of the four assessments will vary across and within each subject. The exceptions will be most mathematics and science subjects, where the external assessment will be weighted at 50 per cent of the total assessment in Units 3 and 4, covering content from across the two units.

External assessments will be developed by teams of subject experts drawn mainly from schools and tertiary institutions. The team of developers will not construct the entire assessment but will instead have responsibility for generating a range of items for possible inclusion in the final assessment task. A 'chief examiner', usually an officer of the QCAA, will be responsible for compiling the various items into a single assessment task, with additional items 'banked' for further refinement and use in subsequent years. The completed assessment will be reviewed by a scrutiny panel, which, alongside double-checking content for instance, will also complete the assessment task in conditions reflecting that expected of students. Trial external assessments have been developed based on existing syllabus requirements, with sample tasks available on the QCAA website. It is intended that indicative external assessment tasks based on the revised syllabi will be developed and distributed in late 2018.

Marking will be undertaken by trained markers who will participate in compulsory calibration activities. Most external assessments will be marked online. Quality assurance of marking will be undertaken through either double-marking or single-marking with additional check marking dependent on the nature of the assessment task (i.e. longer response tasks, such as analytical essays in English, will be double-marked, while short answer questions will be check marked only). Double-marking will involve

the blind re-marking of all scripts by at least two markers. Where there is a discrepancy in the marks, a third referee marker will also mark the script, with a process of mark resolution undertaken. Check marking will only occur for scripts that are single marked, and will involve the sampling of scripts and a review of marking by experienced markers to confirm that the allocation of marks is in line with marking guides. In both processes, control scripts are used to ascertain any markers who are marking outside variance allowances. Where this occurs, a process of recalibration will be undertaken and, where necessary, scripts re-marked. This model reflects current QCAA practice in marking the QCS Test. It is intended that a period of two weeks will be required for marking, with an additional week available for re-marking or reviews of scripts.

School-based assessment (coursework)

For general subjects, each student will complete three formal school-based assessments (in addition to an external assessment) to meet certification requirements. In applied subjects, all assessment will be school-based. For two applied subjects, Essential English and Essential Mathematics, one school-based assessment will be a common task developed by the QCAA, but implemented and marked by individual schools using a common marking scheme developed by the QCAA. The school-based assessment requirements are described within the syllabus with guidelines for teachers on the conditions and techniques for assessment.

Reliability and comparability of school-based assessment results will be supported through processes of endorsement and confirmation. Endorsement of school-based assessment will occur prior to teaching of the content, with the school required to present to the QCAA proposed assessment tasks and detailed student expectations so that they may be reviewed and endorsed. The syllabus documents mandate particular assessment approaches (e.g. prescribing in chemistry that a student should complete a written data test and a first-hand experimental investigation), but the syllabi allow teachers to contextualize assessments to the particular characteristics of the school and students. An example of syllabus guidelines is provided in Annex 1.

School-based assessment will be marked by the classroom teacher, using instrument specific marking guides (ISMGs) provided in the syllabus (see example in Annex 1). The ISMGs provide a structure where judgements are able to be made against the criteria of the task and the expected standards of the syllabus. All marks will be provided to the QCAA prior to the commencement of the confirmation process. Confirmation will involve

the sampling of student work across a range of achievement with individual samples being determined by the QCAA. A network of assessors will review the student work against the prescribed marking criteria (based on the ISMGs), confirming the accuracy of the result awarded by the classroom teacher. These assessors will undergo a process of formal training and calibration activities to ensure consistency of judgements. Student results may be adjusted to reflect variation in assessors' judgements, with the exact policy and practice to be formulated.

Standard setting process

Determining grades

Syllabi within the new senior system outline the rationale, content, assessments and marking guides for each subject. This has signalled a move in Queensland towards higher-definition syllabi that provide greater guidance for teachers in designing curriculum and assessing student achievement. Current syllabi provide broad guidance for teachers, from which more detailed work plans are developed, allowing greater flexibility and accounting for diverse teaching contexts (Luke *et al*., 2008). The new syllabi provide greater prescription of curriculum content and assessment, which should be expected to have an impact on pedagogical practices (Menter and Hulme, 2013). Most significantly, the number of required summative assessments has been greatly reduced with the intent to make more time available for focus on teaching. However, teacher professional judgement will continue to play a significant role in the assessment and determination of student grades and outcomes.

Queensland teachers have a long history of reporting student achievement based on evidence that they have collected from school-based assessment. This is an important consequence of valuing different techniques of assessment and seeking to provide teachers with professional development. The ISMGs in the syllabus documents describe the expected qualities of student work and can be used to discuss the quality of individual student responses during the marking, moderation and confirmation processes. The marking guides reflect the expected standards of student achievement developed with reference to a hierarchy of cognitions based on the work of Marzano and Kendall (2007). This approach ensures that teachers are still able to contextualize their assessment, but student outcomes and achievements are easily compared across different settings.

Final subject results for general subjects will be derived from a combination of the three school-based assessments and one external assessment. The results across the four assessment tasks will not be scaled

against each other but will instead be combined to provide an overall result. In this way, the assessment decisions of teachers will not be subordinate to the results from external assessments.

It is intended that the combined internal and external raw scores will be mapped to a scale of related syllabus standards using a method known as 'item-descriptor matching methods' (Cizek and Bunch, 2007). A modified Rasch model analysis will be used to establish standards cut scores, with inputs coming from the individual marks or grades awarded according to the specific items or criteria contained in the ISMGs for school-based assessment, and criteria of the external assessment. Verification of the standards cut scores will be undertaken through sampling and review of borderline student samples. A panel of expert assessors and reviewers will review student work where the results are close to the proposed boundary of scores for a particular standard to consider the suitability of the application of the standards.

Final results in general subjects will be reported to students as a numerical result out of 100, with achievement of standards presented on an A–E scale, where a C standard is equivalent to a student of satisfactory achievement of the expected standards of learning (Department of Education and Training [Qld], 2016b). For applied subjects, only the A–E grade will be reported. Applied subjects will not have external assessment, and a student may only use one applied subject in the calculation of a tertiary entrance rank. The reported marks for general subjects will be the combined raw scores across the school-based and external assessments. It is expected, though not controlled for, that the form and expectations of school-based assessment will not vary significantly from year-to-year. The stability of parameters within the syllabi, and the capacity of the endorsement process to ensure assessment of comparable difficulty, should allow for comparisons across year groups to occur for the purpose of deriving cut scores, and confidence that achievement of a particular standard in one year is comparable to achievement of the same standard in another year. Calculation of an ATAR will be a separate process which ranks student overall achievement and is not standards based, but simply employs the final results in a process of calculations.

Political and public controversies/debates with the Queensland Certificate of Education

The QCE was introduced in 2008 and is therefore only a relatively recent qualification, although the current approach to assessment in Queensland has a rich history. The focus of recent concerns or debates has not been in

relation to the qualification itself but rather the reliance on school-based assessment for determining student achievement and its relationship to tertiary entrance and preparedness for future study.

It is tempting to represent the impending renewal of senior assessment as the result of concerns about the reliability of school-based assessment as it is currently implemented in Queensland. Despite over four decades of experience, there have always been criticisms of the system from those committed to the use of subject-based external assessment (Allen, 2013). However, it is argued here that the motivation for curriculum renewal is more accurately attributed to systemic changes in the broader educational environment that have occurred over the past two decades (McCulloch, 1998; Sinnema and Aitken, 2013). These changes have had a significant impact on the interface between secondary and tertiary education.

Some of the most influential changes that have impacted on the secondary schooling sphere are:

- the changing nature of schooling: Greater numbers of students are now completing Year 12, with more of these students seeking to continue studying after school. This has impacted on the purpose of senior schooling and its expected outcomes
- the blurring of boundaries between secondary and tertiary education: There has been an increased uptake of vocational education and training during the senior phase of schooling, and significant numbers of students studying subjects at university while they are still at school
- a new and more flexible senior qualification: The QCE was introduced to recognize and encourage a wide range of learning options and impose minimum standards of literacy and numeracy
- the increasing influence of the Global Education Reform Movement (GERM): The emergence of assumptions that educational improvements come from competition and accountability, such as ranking schools based on common national assessment results, standards-based assessment and prescriptive and homogeneous curricula focused on literacy, numeracy and knowledge and skills in science (Sahlberg, 2011). There has been an associated movement in the foundation of educational policy towards achievement on international measures and global competitiveness
- publication of student data and school comparison tables: The official publication of student achievement data and increasing tendency for media outlets to use both official and unofficial data to generate league tables purporting to compare achievement between schools has led

to schools focusing on achieving reportable outcomes for students. One possible result has been an increase in student preparation for the annual QCS Test at the expense of subject-based teaching and assessment.

Mainly occurring at the national level, other challenges have arisen from changes in the tertiary sector and beyond:

- movement towards a common tertiary entrance rank: All states and territories, other than Queensland, embraced a common ATAR from 2009
- deregulation of the higher education sector: Following the Review of Australian Higher Education (Bradley *et al*., 2008), new participation targets were set for Australian universities, and the deregulation of previous quota restrictions on university places was introduced. These changes resulted in increased competition between tertiary institutions and an imperative to attract students in greater numbers. This has led tertiary providers to use alternatives to senior certification and ranking to offer direct entry to Year 12 students. Recent reports indicate that only 31 per cent of students are admitted to universities based solely on their ATAR (HESP, 2016)
- fluctuations and changes in the employment market caused by economic conditions: The Queensland and Australian economies have been significantly affected by a range of changing conditions including the recent global financial crisis and mining booms and contractions. These changing conditions have created varying demands for particular skills, which have impacted on the value associated with particular areas of study.

While the systemic changes taking place in Queensland are best understood as the consequence of a wide range of factors, the parliamentary inquiry into assessment methods used in mathematics and science subjects, and the ACER review of senior assessment and tertiary entrance processes, both demonstrated that there was also a divide between policy and practice regarding the purpose of assessment, and the use and reporting of standards in the Queensland system. For example, critics of the system frequently expressed concerns that were based on an assumption that marks were not allowed to be used in a standards-based system and disagreed with the concept of a student being able to demonstrate different levels of achievement in different questions or tasks (Queensland Parliament, 2013c). There were also perceptions that the moderation system was not sufficiently robust to

deliver fair and accurate outcomes, and that it should only continue with the addition of external examinations used for scaling purposes (Matters and Masters, 2014).

The inquiry concluded that the doubt about whether it is possible to make valid and reliable judgements of student achievement in senior mathematics, chemistry, and physics using only school-based assessment, related in part to:

- teachers' lack of support for the assessment methods of these particular syllabi (which had undergone significant change in the previous decade)
- a lack of a common assessment that allows direct comparison of students (Queensland Parliament, 2013c).

These views challenged the body of research that has demonstrated how social moderation of student assessment supports the ongoing professional learning of teachers and can be as reliable as external examinations (see, for example, Hipkins and Robertson, 2011; Klenowski and Wyatt-Smith, 2010). As the ACER reviewers asserted in reference to school-based assessments, 'The reliability and comparability of such assessments depend in part on the assessment activities themselves. In general, the more tightly specified and similar the activities on which assessments are made, the more reliable and comparable the resulting judgements' (Matters and Masters, 2014: 48). It was this conclusion that led ACER to propose, and for the government to accept, the new processes of endorsement and confirmation mentioned above. Importantly, the new Queensland system goes further than just incorporating external assessment by introducing new approaches for the state-wide endorsement of school-based assessment tasks before they are used in the classroom and the provision of professional learning and accreditation of assessors within the system. Both of these approaches are designed to further enhance the quality of assessment tasks, better supporting the fair and reliable application of standards of achievement to student responses to assessment tasks.

ACER's recommendations reinforce the view that the focus of a functional assessment system should be on assessment quality and its validity or fitness-for-purpose. The starting point in designing an assessment programme should be to identify the total body of evidence needed to judge student achievement. If each is understood to be inherently valid, it is possible for school-based and external assessment to coexist constructively. The greater consistency and transparency of external assessments can be effectively combined with the more familiar deep learning and engagement produced by school-based assessments that include projects, reports,

investigations, orals, practical work, fieldwork, performances, presentations, essays, examinations and the production of artefacts.

Annex 1: Sample syllabus assessment and marking guide (Chemistry)

The following example is indicative only, with final approval of syllabus content and assessment requirements still to be provided.

Description

This assessment requires students to research a question or hypothesis through collection, analysis and synthesis of primary data. A student experiment uses investigative practices to assess a range of cognitions in a particular context. Investigative practices include locating and using information beyond students' own knowledge and the data they have been given.

Research conventions must be adhered to. This assessment occurs over an extended and defined period of time. Students may use class time and their own time to develop a response.

Assessment objectives

This assessment technique is used to determine student achievement in the following objectives (note that Objective 1 is not assessed in this instrument):

2. *apply understanding* of chemical equilibrium systems and oxidation and reduction to *modify experimental* methodologies and process primary *data*
3. *analyse experimental evidence* about chemical equilibrium systems or oxidation and reduction
4. *interpret experimental evidence* about chemical equilibrium systems or oxidation and reduction
5. *investigate* chemical equilibrium systems or oxidation and reduction through an *experiment*
6. *evaluate experimental processes* and *conclusions* about chemical equilibrium systems or oxidation and reduction
7. *communicate understandings* and *experimental findings*, *arguments* and *conclusions* about chemical equilibrium systems or oxidation and reduction.

Specifications

Description

In the student experiment, students *modify* (i.e. refine, extend, or redirect) an *experiment* in order to address their own related *hypothesis* or question. It is *sufficient* that students use a practical performed in class or a *simulation* as the basis for their *methodology* and *research question*.

In order to complete the assessment task, students must (note that the steps indicated with an asterisk * below may be completed in groups. All the other elements must be completed individually):

- identify an experiment to modify*
- develop a research question to be investigated*
- research relevant background scientific information to inform the modification of the research question and methodology
- conduct a risk assessment and account for risks in the methodology*
- conduct the experiment*
- collect sufficient and relevant qualitative and/or quantitative data to address the research question*
- process and present the data appropriately
- analyse the evidence to identify trends, patterns, or relationships
- analyse the evidence to identify uncertainty and limitations
- interpret the evidence to draw conclusion/s to the research question
- evaluate the reliability and validity of the experimental process
- suggest possible improvements and extensions to the experiment
- communicate findings in an appropriate scientific genre (e.g. poster, report, journal article, conference presentation).

Scientific inquiry is a non-linear, iterative process. Students will not necessarily complete these steps in the stated order; some steps may be repeated or revisited.

Conditions

- Time: 10 hours class time. This time will not necessarily be sequential. Students must perform the majority of the tasks during class time, including
  - performing background research and developing the methodology
  - conducting the experiment

- processing and analysing evidence and evaluating the methodology
- preparing and presenting the response (e.g. writing the report, constructing and presenting the poster).

- Length:
  - written (e.g. scientific report), 1,500–2,000 words
    or
  - multimodal presentation (e.g. poster presentation), 9–11 minutes.

- Other:
  - students may work collaboratively with other students to develop the methodology and perform the experiment; all other stages (e.g. processing of data, analysis of evidence, and evaluation of the experimental process) must be carried out individually
  - the response must be presented using an appropriate scientific genre (e.g. scientific report, poster presentation, logbook entries, conference presentation) and contain
    - a research question
    - a rationale for the experiment
    - reference to the initial experiment and identification and justification of modifications to the methodology
    - raw and processed qualitative and/or quantitative data
    - analysis of the evidence
    - conclusion/s based on the interpretation of the evidence
    - evaluation of the methodology and suggestions of improvements and extensions to the experiment
    - a reference list

Instrument-specific marking guide (indicative only)

Criterion: Research and planning

ASSESSMENT OBJECTIVES

2. apply understanding of chemical equilibrium systems or oxidation and reduction to modify experimental methodologies and process primary data
5. investigate chemical equilibrium systems or oxidation and reduction through an experiment

Table 10.1: The characteristics of the student work

Characteristics	Marks
• informed application of understanding of chemical equilibrium systems or oxidation and reduction to modify experimental methodologies demonstrated by – a considered rationale for the experiment – justified modifications to the methodology • effective and efficient investigation of chemical equilibrium systems or oxidation and reduction demonstrated by – a specific and relevant research question – a considered methodology that enables the collection of sufficient, relevant data – considered management of risks and ethical or environmental issues.	5–6
• adequate application of understanding of chemical equilibrium systems or oxidation and reduction to modify experimental methodologies demonstrated by – a reasonable rationale for the experiment – feasible modifications to the methodology • effective investigation of chemical equilibrium systems or oxidation and reduction demonstrated by – a relevant research question – a methodology that enables the collection of relevant data – management of risks and ethical or environmental issues.	3–4
• rudimentary application of chemical equilibrium systems or oxidation and reduction demonstrated by – a vague or irrelevant rationale for the experiment – inappropriate modifications to the methodology • ineffective investigation of chemical equilibrium systems or oxidation and reduction demonstrated by – an inappropriate research question – a methodology that causes the collection of insufficient and irrelevant data – inadequate management of risks and ethical or environmental issues.	1–2
• does not satisfy any of the descriptors above.	0

Table 10.2: The characteristics of the student work

Characteristics	Marks
• appropriate application of algorithms, visual and graphical representations of data about chemical equilibrium systems or oxidation and reduction demonstrated by correct and relevant processing of data • systematic and effective analysis of experimental evidence about chemical equilibrium systems or oxidation and reduction demonstrated by – thorough identification of relevant trends, patterns or relationships – thorough and appropriate identification of the uncertainty and limitations of the evidence • effective and efficient investigation of chemical equilibrium systems or oxidation and reduction demonstrated by the collection of sufficient and relevant raw data.	5–6
• adequate application of algorithms, visual and graphical representations of data about chemical equilibrium systems or oxidation and reduction demonstrated by basic processing of data • effective analysis of experimental evidence about chemical equilibrium systems or oxidation and reduction demonstrated by – identification of obvious trends, patterns or relationships – basic identification of uncertainty and limitations of evidence • effective investigation of chemical equilibrium systems or oxidation and reduction demonstrated by the collection of relevant raw data.	3–4
• rudimentary application of algorithms, visual and graphical representations of data about chemical equilibrium systems or oxidation and reduction demonstrated by incorrect or irrelevant processing of data • ineffective analysis of evidence demonstrated by – identification of incorrect or irrelevant trends, patterns or relationships – incorrect or insufficient identification of uncertainty and limitations of evidence • ineffective investigation of chemical equilibrium systems or oxidation and reduction demonstrated by the collection of insufficient and irrelevant raw data.	1–2
• does not satisfy any of the descriptors above.	0

Table 10.3: The characteristics of the student work

Characteristics	Marks
• insightful interpretation of experimental evidence about chemical equilibrium systems or oxidation and reduction demonstrated by justified conclusion/s linked to the research question • critical evaluation of experimental processes about chemical equilibrium systems or oxidation and reduction demonstrated by – justified discussion of the reliability and validity of the experimental process – suggested improvements and extensions to the experiment which are logically derived from the analysis of the evidence.	5–6
• adequate interpretation of experimental evidence about chemical equilibrium systems or oxidation and reduction demonstrated by reasonable conclusion/s relevant to the research question • basic evaluation of experimental processes about chemical equilibrium systems or oxidation and reduction demonstrated by – reasonable description of the reliability and validity of the experimental process – suggested improvements and extensions to the experiment which are related to the analysis of the evidence.	3–4
• invalid interpretation of experimental evidence about chemical equilibrium systems or oxidation and reduction demonstrated by inappropriate or irrelevant conclusion/s • superficial evaluation of experimental processes about chemical equilibrium systems or oxidation and reduction demonstrated by – cursory or simplistic statements about the reliability and validity of the experimental process – ineffective or irrelevant suggestions.	1–2
• does not satisfy any of the descriptors above.	0

Criterion: Analysis of evidence

ASSESSMENT OBJECTIVES

2. apply understanding of chemical equilibrium systems or oxidation and reduction to modify experimental methodologies and process primary data

3. analyse experimental evidence about chemical equilibrium systems or oxidation and reduction
5. investigate chemical equilibrium systems or oxidation and reduction through an experiment

Criterion: Interpretation and evaluation

ASSESSMENT OBJECTIVES

4. interpret experimental evidence about chemical equilibrium systems or oxidation and reduction
6. evaluate experimental processes and conclusions about chemical equilibrium systems or oxidation and reduction

Criterion: Communication

Assessment objective

7. communicate understandings and experimental findings, arguments and conclusions about chemical equilibrium systems or oxidation and reduction

Table 10.4: The characteristics of the student work

Characteristics	Marks
• effective communication of understandings, findings, arguments and conclusions about chemical equilibrium systems or oxidation and reduction demonstrated by – fluent and concise use of scientific language and representations – appropriate use of genre conventions – acknowledgment of sources of information through appropriate use of referencing conventions.	2
• adequate communication of understandings, findings, arguments and conclusions about chemical equilibrium systems or oxidation and reduction demonstrated by – competent use of scientific language and representations – use of basic genre conventions – use of basic referencing conventions.	1
• does not satisfy any of the descriptors above.	0

References

Allen, R. (2013) *Strengths and Weaknesses of Queensland's OP System Today*. Melbourne: Australian Council for Educational Research. Online. https://goo.gl/uMrPKY (accessed 12 June 2018).

Blyth, K. (2014) 'Selection methods for undergraduate admissions in Australia: Does the Australian predominate entry scheme the Australian Tertiary Admissions Rank (ATAR) have a future?'. *Journal of Higher Education Policy and Management*, 36 (3), 268–78.

Bradley, D., Noonan, P., Nugent, H. and Scales, B. (2008) *Review of Australian Higher Education: Final report*. Canberra: Department of Education, Employment and Workplace Relations. Online. http://hdl.voced.edu.au/10707/44384 (accessed 12 June 2018).

Clarke, E. (1987) *Assessment in Queensland Secondary Schools: Two decades of change 1964–1983*. Brisbane: Department of Education. https://goo.gl/V3Poij (accessed 12 June 2018).

Cizek, G.J. and Bunch, M.B. (2007) *Standard Setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.

Department of Education and Training (Cwth) (2015) *Undergraduate Applications, Offers and Acceptances 2015*. Canberra: Department of Education and Training, Australian Government (Cwth). Online. https://goo.gl/3Azd9o (accessed 12 June 2018).

Department of Education and Training (Qld) (2016a) *Statistics and Information*. Brisbane: Department of Education and Training, Queensland Government. Online. https://qed.qld.gov.au/publications/reports/statistics/schooling (accessed 12 June 2018).

Department of Education and Training (Qld) (2016b) *Reporting to Parents*. Brisbane: Department of Education and Training, Queensland Government. Online. http://education.qld.gov.au/curriculum/framework/p-12/docs/policy-reporting.doc (accessed 12 June 2018).

HESP (Higher Education Standards Panel) (2016) *Improving the Transparency of Higher Education Admissions*. Canberra: Department of Education and Training. Online. https://goo.gl/SBxqHb (accessed 12 June 2018).

Hipkins, R. and Robertson, S. (2011) *Moderation and Teacher Learning: What can research tell us about their interrelationships?* Wellington: New Zealand Council for Educational Research. Online. http://www.nzcer.org.nz/system/files/moderation-teacher-learning.pdf (accessed 12 June 2018).

Kelly, D. (2014) *An Analysis of Earlier Reports into Senior Assessment and Tertiary Entrance Procedures in Queensland*. Melbourne: Australian Council for Educational Research. Online. www.acer.edu.au/files/Analysis_of_reports.pdf (accessed 12 June 2018).

Klenowski, V. and Wyatt-Smith, C. (2010) 'Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform'. *Assessment Matters*, 2, 107–131.

Luke, A., Weir, K. and Woods, A. (2008) *Development of a Set of Principles to Guide a P-12 Syllabus Framework: A report to the Queensland Studies Authority*. Brisbane: Queensland Government. Online. https://goo.gl/pmvvTt (accessed 12 June 2018).

Matters, G. and Masters, G. (2014) *Redesigning the Secondary–Tertiary Interface: Queensland review of senior assessment and tertiary entrance*. Brisbane: Australian Council for Educational Research. Online. https://goo.gl/UFK7fy (accessed 12 June 2018).

Marzano, R.J. and Kendall, J.S. (2007) *The New Taxonomy of Educational Objectives*. 2nd ed. Thousand Oaks, CA: Corwin Press.

McCulloch, G. (1998) 'Curriculum Reform, Educational Change and School Improvement'. In Hargreaves, A., Lieberman, A., Fullan, M., Hopkins, D. (eds) *International Handbook of Educational Change*. Kluwer International Handbooks of Education, vol 5. Dordrecht: Springer, 1203–15.

Menter, I. and Hulme, M. (2013) 'Developing the teacher – or not?'. In Priestley, M. and Biesta, G. (eds) *Reinventing the Curriculum: New trends in curriculum policy & practice*. London: Bloomsbury Academic, 132–48.

Pitman, J.A. (1987) *Tertiary Entrance in Queensland: A review*. Brisbane: Board of Secondary School Studies.

QCEC (Queensland Catholic Education Commission) (2016) *Catholic Education*. Brisbane: Queensland Catholic Education Commission. Online. http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/4221.02017?OpenDocument (accessed 27 September 2016).

QCAA (Queensland Curriculum and Assessment Authority) (2014) *Education Act 2014*. Brisbane: Queensland Government.

Queensland Parliament (2013a) *The Assessment Methods Used in Senior Mathematics, Chemistry and Physics in Queensland Schools*. Parliamentary Committees. Education and Innovation Committee Report No. 25. Online. https://goo.gl/qUPVmk (accessed 13 June 2018).

Queensland Parliament (2013b) *Queensland Government Response*. Education and Innovation Committee Report No. 25. Online. https://goo.gl/FWqqFD (accessed 13 June 2018).

Queensland Parliament (2013c) *Education and Innovation Committee Public Briefing Inquiry into the Assessment Methods Used in Senior Mathematics, Chemistry and Physics in Queensland Schools*. Online. https://goo.gl/6RZrh2 (accessed 17 June 2018).

QGSO (Queensland Government Statisticians Office) (2011) *Census 2011: Education in Queensland*. Brisbane: Queensland Government Statistical Office. Online. https://goo.gl/gia5Hv (accessed 26 September 2016).

Radford, W.C. (1970) *Public Examinations for Queensland Secondary School Students: Report of the committee appointed to review the system of public examinations for Queensland secondary school students and to make recommendations for the assessment of students achievements*. Brisbane: Department of Education and Training (Qld). Online. https://goo.gl/F8v1kz (accessed 17 June 2018).

Sahlberg, P. (2011) *Finnish Lessons: What can the world learn from educational change in Finland?* New York: Teachers College Press.

Scott, D.R., Berkeley, G.F., Howell, M.A., Schuntner, L.T., Walker, R.F. and Winkle, L. (1978) *A Review of School-Based Assessment in Queensland Secondary Schools*. Brisbane: Board of Secondary School Studies. Online. https://goo.gl/fwc23W (accessed 17 June 2018).

Sinnema, C. and Aitken, G. (2013) 'Emerging international trends in curriculum'. In Priestley, M. and Biesta, G. (eds) *Reinventing the Curriculum: New trends in curriculum policy & practice*. London: Bloomsbury Academic, 114–31.

Viviani, N. (1990) *The Review of Tertiary Entrance in Queensland 1990*. Brisbane: Department of Education. Online. https://goo.gl/Q1S9rn (accessed 17 June 2018).

Managing the tension between performance standards and aggregate ranking

Graham S. Maxwell

After almost 50 years of successful and robust senior secondary school certification based solely on school-based assessment (Maxwell and Cumming, 2011), Queensland has finally succumbed to pressures for greater alignment with other Australian states and territories. Matthew Campbell has done well in describing some of the features of the old and new systems. It is important to note the distinction between certification of a performance standard (or level of achievement) in each subject studied by a student, and the combination of each student's results across several subjects to produce a rank ordering of all students for purposes of university (tertiary) admission. In Queensland, these two will continue to be kept conceptually and operationally separate, and it is desirable to do so, since they are two quite different measures with different meanings (certification of subject performance against explicit performance standards versus relative ranking of general performance summed across whichever subjects were studied).

There is some consistency between old and new in Queensland. Subject achievement will continue to be reported in terms of five expressed standards, and school-based assessments will be socially moderated (by external 'verifiers' rather than moderation panels). Moderation through teacher professional judgement (Maxwell, 2010) remains at the heart of the within-school and between-school processes. The new procedures are, however, more prescriptive, and some of the previous flexibility for adaptation to local and individual circumstances would seem to have been lost; for example, it is unclear how accommodations for special needs and 'make-ups' for illness, etc., will be managed, something previously decided within the context of the school without the need for external approval. The new endorsement and confirmation processes will be much more centrally controlled than the old approval and confirmation processes, which were much more advisory and negotiable.

A major difference will be the way in which a subject result is determined. The old approach involved holistic teacher judgement of a portfolio of assessments against subject performance standards, with those judgements being the focus of the moderation procedures. The new approach focuses on verification (moderation) of each of the school-based assessments, involving professional judgement using 'item specific marking guides'. The school-based results and the external assessment results are to be numerically combined and cut-offs for the subject grades established on this combined scale, although professional judgement remains at the heart of this process. Reporting subject results on a 100-point scale, not just the grades, is a major change; previously, any accompanying numerical results were not reported.

There has clearly been considerable thought given to the new system. No doubt fine-tuning will be needed over time. Research into comparative qualities and effects of the old and new systems ought to be a priority.

The calculation of a rank ordering of all students completing a senior secondary school programme of study that makes them eligible for university studies is a peculiarly Australian practice. In other places, a grade point average would suffice. Why then the calculation of the Australian Tertiary Admission Rank (ATAR)? The answer would seem to lie in the diversity of senior secondary school subjects on the one hand and diversity of university undergraduate courses on the other.

First, ATAR attempts to adjust for differences in the quality of the cohorts of students choosing to study different subjects. It does this by scaling subject results psychometrically against a measure of 'overall achievement'. In the current Queensland system, a similar measure, the Overall Position (OP), is derived by scaling subject results against the Queensland Core Skills Test as the moderator measure of overall achievement. For ATAR, practice across the states is inconsistent but essentially an iterative other-subject scaling process (sum each student's several subject results, use the resulting measure to moderate each subject result, and repeat, preferably until there is convergence). ATAR is expressed as a percentile with increments of 0.05. The current OPs in Queensland are reported on 25 ranks, based on data simulations showing finer distinctions were unwarranted. The new procedures in Queensland can presumably deliver greater precision. However, in general, ATARs would appear to be expressed at an unwarranted level of precision (essentially a 2,000-point scale).

Second, in Australia, with some exceptions, university admission is mostly to a particular undergraduate specialization, not to the university as a whole. This requires a rapid sorting of offers and acceptances, and

universities have established the system of state tertiary admissions centres to do this. Applicants receive a single offer based on their course preferences. Apart from some specialist performance areas such as music, universities do not in general select students based on transcripts, portfolios, presentations or interviews. Instead, each state has a Tertiary Entrance Centre that sorts student preferences based on the ATAR and makes a single offer of a university place. A great deal therefore hangs on the ATAR as a competitive ranking of student quality. In most cases, there is no set standard for entry to a university course; cut-offs for entry depend on the competition for available places.

References

Maxwell, G.S. (2010) 'Moderation of student work by teachers'. In McGaw, B., Baker, E. and Peterson, P. (eds) *International Encyclopedia of Education* (Vol. 3). Oxford: Elsevier, 457–63.

Maxwell, G.S. and Cumming, J.J. (2011) 'Managing without public examinations: Successful and sustained curriculum and assessment reform in Queensland'. In Yates, L., Collins, C. and O'Connor, K. (eds) *Australia's Curriculum Dilemmas: State perspectives and changing times*. Melbourne: Melbourne University Press, 202–22.

The curious case of Queensland and a middle way for senior schooling assessment

Joshua McGrane

Campbell's chapter highlights the historically unique approach of Queensland to senior schooling and tertiary entrance assessment in the Australian context, particularly in the sole reliance upon school-based assessments for these high stakes purposes. While other Australian jurisdictions have also historically had their own standards-referenced assessment and external moderation practices for senior schooling assessment (Wyatt-Smith *et al*., 2017), external examinations have typically taken precedence. This precedence is reflected in the use of external examinations to moderate (or 'scale') the school-based marks, as these examinations are perceived as more reliable and trustworthy, even though they are potentially limited in terms of providing more contextualized and authentic assessment of students' learning (Maxwell, 2006).

The recent reforms in Queensland, including the reintroduction of external examinations, bring their practices closer to other Australian jurisdictions. Nonetheless, the continuing systemic emphasis upon teacher-driven assessment in this high stakes context, which is blended with improved centralized control and monitoring processes to ensure the reliability and comparability of these assessments, represents a middle way in senior schooling assessment. This middle way balances a trust in teachers' professionalism and assessment practices, underpinned by an emphasis upon assessment-related professional development, with centralized processes concerned with accountability and gathering evidence to ensure that the individual schools' and teachers' assessment practices reflect the systemic and curricular expectations (Hopfenbeck *et al*., 2015). As a result, the success (or otherwise) of the reformed Queensland senior schooling assessment system will be an interesting case study for researchers and policymakers interested in a balanced approach to high stakes educational assessment.

Queensland's teacher-centred approach to senior-school assessment is consistent with a more general push in Australia to train assessment-capable teachers, as reflected in the national standards for teaching and teacher training introduced in 2012 (Wyatt-Smith *et al*., 2017). Despite this push, there is a scarcity of published research on the reliability and validity of high stakes teacher assessments in Australia. Johnson (2013) suggests that this scarcity is reflected globally, and the limited evidence available on this topic suggests that teachers are commonly influenced by construct-irrelevant factors when making their assessments, including the gender, socio-economic background, effort and behaviour of their students. Hay and Macdonald's (2008) case-study of senior secondary Physical Education teachers in the Queensland context was consistent with this claim. Teachers were found to make their judgements along somewhat 'intuitive' lines and were influenced by the attitudes and sporting histories of the students, rather than explicitly referenced to the criteria and standards of student performances set out in the syllabus. Based on this limited evidence, the reforms to the Queensland assessment system to provide additional oversight, scaffolding and resources to teachers to ensure that their assessments are explicitly referenced to the relevant syllabus are welcome. Nonetheless, the influence of construct-irrelevant factors on the teacher assessments should be a key concern in monitoring processes and explicitly addressed in teacher training and professional development in assessment.

On a more critical note, the requirement for teachers to assess students' performances by rating them on a small number of coarse-grained levels, along with the use of marking guides that are somewhat generic with respect to the specific assessment tasks, are concerning elements of Queensland's reformed assessment system. Andrich (2006) argued that similar features present in the Western Australian assessment system at the time led to systematic biases in teacher judgements and were insufficiently precise for purposes of tertiary entrance selection. Therefore, as the Queensland system is further developed and implemented, an eye should be kept to the mistakes made by other Australian jurisdictions in the past.

References

Andrich, D. (2006) *A Report to the Curriculum Council of Western Australia Regarding Assessment for Tertiary Selection* (Andrich Report). Online. https://goo.gl/ZDn52g (accessed 17 June 2018).

Hay, P.J. and Macdonald, D. (2008) '(Mis)appropriations of criteria and standards-referenced assessment in a performance-based subject'. *Assessment in Education: Principles, policy & practice*, 15 (2), 153–68.

Hopfenbeck, T.N., Flórez Petour, M.T. and Tolo, A. (2015) 'Balancing tensions in educational policy reforms: Large-scale implementation of Assessment for Learning in Norway'. *Assessment in Education: Principles, Policy & Practice*, 22 (1), 44–60.

Johnson, S. (2013) 'On the reliability of high-stakes teacher assessment'. *Research Papers in Education*, 28 (1), 91–105.

Maxwell, G. (2006) 'Quality management of school-based assessments: Moderation of teacher judgments'. Paper presented at the 32nd International Association for Educational Assessment (IAEA) Conference, Singapore, May. Online. https://goo.gl/JQfwB7 (accessed 17 June 2018).

Wyatt-Smith, C., Alexander, C., Fishburn, D. and McMahon, P. (2017) 'Standards of practice to standards of evidence: Developing assessment capable teachers'. *Assessment in Education: Principles, Policy & Practice*, 24 (2), 250–70.

Chapter 11

Standard Setting in South Africa: The National Senior Certificate

Emmanuel Sibanda

Introduction

South Africa is a country on the southernmost tip of the African continent. The total population in South Africa is estimated at 55.6 million people, according to the latest Community Survey 2016 figures (StatSA, 2016). Since 1994, South Africa has been divided into nine provinces. They vary widely in population, from the mostly urban Gauteng, which contains over 20 per cent of the national population, to the mostly desert Northern Cape, which contains less than 3 per cent. Other provinces are KwaZulu-Natal (19.9 per cent); Eastern Cape (12.6 per cent); Western Cape (11.3 per cent); Limpopo (10.4 per cent); Mpumalanga (7.8 per cent); North West (6.7 per cent); and Free State (5.1 per cent).

Overview of the South African education system

In South Africa, the Department of Basic Education (DBE) is responsible for formulating, developing and reviewing policies and legislation in respect of the education system from Grade R to 12. The education system could be regarded as an 8 + 2 + 3 system. The thirteen years of schooling are divided into four phases as follows:

Table 11.1: Organization of primary and secondary schooling in South Africa

Phases	Grades	Schools	Age Range
Foundation	R, 1–3	Primary	6–9
Intermediate	4–7		10–13
Senior	8–9	Senior/High	14–15
Further Education and Training (FET)	10–12		16–18

There are eight years of primary schooling (broken up into Grade R, foundation and intermediate phases), followed by two years of senior/high school, which together make up general education and training, the compulsory component of schooling. This is followed by three years of further education and training. There are no public examinations at the end of grade nine, and no national certificate is issued; learners are issued report cards for each grade by their schools. The public examinations or external examinations take place only at the end of Year 13 or Grade 12. The average ages of Grade 12 cohorts of learners are 17 and 18. As in most countries, Grade 12 examinations serve a dual role: as a school exit and as a portal into tertiary education. According to the DBE in 2016, of the 1.23 million learners who enrolled for grade one in 2005, only 657,447 (53 per cent) registered for the Grade 12 examinations in 2016. Some of the learners continued at vocational colleges and others dropped out from school. The majority of the dropouts occurred in Grades 10 and 11. On average, between 50 per cent and 55 per cent of learners who enrolled in Grade R proceed to Grade 12 after 12 years of schooling (DBE, 2016a).

There is no single examination body in South Africa. Three assessment bodies administer examinations. In the public schooling system, the government ministry, the Department of Basic Education, sets and administers examinations.

In addition to the public system, two independent examination bodies set examinations for independent schools, the Independent Examinations Board (IEB), which services a large number of independent schools, and the South African Comprehensive Assessment Institute (SACAI). All three assessment bodies set papers in two languages, English and Afrikaans.

The IEB and SACAI are accredited by Umalusi to administer the examinations. The DBE, by law, is deemed accredited. The DBE, IEB and SACAI are regarded as assessment bodies, which is different from what is referred to as an examination board in other countries.

Umalusi was established through the promulgation of the General and Further Education and Training Quality Assurance (GENFETQA) Act (58 of 2001), amended in 2008. Among other things, Umalusi is mandated to (1) develop and implement policy for quality assurance of the assessment (assessments at exit points and site-based assessment); and (2) issue certificates to learners who have achieved qualifications.

Umalusi's role in examinations is through the processes of quality assurance of assessment. These processes are:

- moderation of question papers
- moderation and verification of school-based assessment (SBA)
- monitoring of the state of readiness to conduct the examinations
- monitoring and auditing of the selection and appointment of markers
- monitoring of the writing of the examinations
- monitoring and verifications of marking
- standardization of learners' marks.

Umalusi is an independent body even though it is funded by the ministry.

History of examinations in South Africa

The examination system in South Africa dates back 15 decades. The first 136 years of the system were characterized by multiple standards and fragmented, racialized approaches to exams (NECT, 2015: 6). The University of Good Hope conducted the first exams in the nineteenth century.

In 1918, the Joint Matriculation Board (JMB) of universities in South Africa took over from the University of Good Hope and was also responsible for setting standards for the matriculation certificate. The matriculation certificate, which became the gateway to universities and many professional careers, was established as the only school leaving certificate recognized by several foreign bodies.

In 1921, eight new departmental examinations were established under the jurisdiction of JMB as the arbitrator of standards. The JMB's approach to maintaining standards was through the control of syllabi and curricula as well as the moderation of question papers. The JMB had a particular view of standards that related to validity or dependability of the examination. The JMB strove to minimize the variations from one year to the next, or from one subject to another (Umalusi, 2006). This is how the standards were established and maintained.

The JMB, during its existence, wrestled with the decentralization of the examination to provinces. Over time, the situation became worse as the national education system of South Africa consisted of 19 different education departments, which implied 19 different examination systems (Terblanche, 1989). These examination systems were divided on ethnic and racial lines.

According to Trümpelmann (1991), it was abundantly clear in the late 1980s that the decentralization of the examination had aggravated the problems relating to control and standards.

In 1992, the South African Certification Council (SAFCERT) was established and took over from JMB. The mandate of SAFCERT was to centralize the certification processes, oversee the standardization of results of the Senior Certificate (SC) and externally moderate all examination papers. The centralization of the certification process was seen as key in portraying a uniform standard (Trümpelmann, 1991). SAFCERT continued with JMB's approach to standards; this was to be expected since JMB was instrumental in establishing SAFCERT.

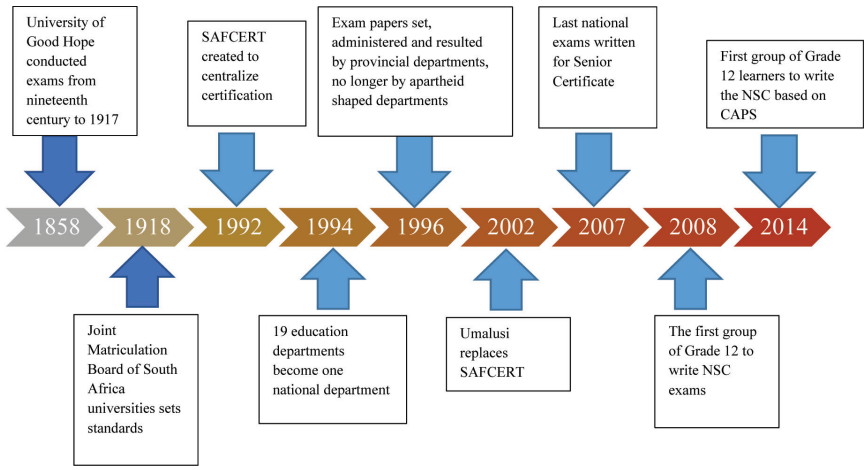
In 1995, the then new government established the first provincial public examination bodies, which came into operation in 1996. The first national examination, under the newly elected democratic government, was administered in November 1996, following a highly decentralized approach (DBE, 2016).

Umalusi, which took over the responsibilities of SAFCERT, was established in 2001 and took over the examination responsibility in 2002. By this time, the SCE was the responsibility of the newly recognized non-racial provincial sub-departments of the Department of Education, lately DBE. However, each of the nine provincial departments continued to be responsible for the setting of their own examination papers. This setting of examination papers by different provincial departments created a challenge for the equivalence of examinations standards across provinces. As a result, in 2000, five subjects with high enrolments were set nationally for the purposes of ensuring common examination standards. By the end of 2007, 11 SCE subjects with high enrolments were set nationally by the Department of Education.

The introduction of the National Senior Certificate (NSC), which was examined in grade 12 for the first time in 2008, was a significant milestone of the new government. Since the end of 2008, all grade-12 learners, irrespective of their race or location, have written the same examination, set by a single national department of education. This implies that all learners are examined on the same standard.

Table 11.2 shows a time line that indicates key changes since 1858.

Table 11.2: Organization of primary and secondary schooling in South Africa



Source: Adapted from NECT 2015: Perspective and lessons on public systems improvement

The NSC was first introduced in 2005 in Grade 10 by the then Minister of Education. It has since been amended quite a number of times with the latest being in 2011 (National Planning Commission, 2011). The amendments were to be expected since this is the first examination that is written by all learners irrespective of race. The NSC is a 130-credit certificate at level 4 on the National Qualifications Framework (NQF). In order for the learner to obtain an NSC, he/she must offer/register seven subjects including three compulsory subjects, namely, home language, life orientation, mathematics or mathematical literacy.

Table 11.3: The National Senior Certificate scale ratings with descriptions

Rating code	Description of competence	Percentage of marks
7	Outstanding	80–100
6	Meritorious	70–79
5	Substantial	60–69
4	Adequate	50–59
3	Moderate	40–49
2	Elementary	30–39
1	Not Achieved	0–29

There are concessions for immigrants and students who experience barriers to learning, such as those who are deaf or dyslexic. The concessions will not be discussed in this report.

The achievements of learners in subjects are reported in a seven-level scale. The scale is fixed from one year to the next. Table 11.3 shows a seven-level scale rating with descriptions.

The Grade 12 examinations serve a dual purpose: as a school exit and as a portal into higher education. For a school exit, which is the lowest pass, a learner must have:

- an achievement rating of 3 in three subjects, one of which is an official language at Home Language level
- an achievement rating of 2 in three subjects
- an achievement rating of 1 in the fourth subject.

The admission requirements for entrance to higher education programmes are set out in terms of section 3 the 1997 Higher Education Act (Act No.101 of 1997). Umalusi indicates on the certificate whether a candidate is qualified to enrol for a Higher Certificate, Diploma or a Bachelor degree at a higher education institution. It is, however, still the prerogative of higher education institutions to set specific admission requirements to particular programmes.

Table 11.4 provides a summary of the key features (or 'rules of combination') of the NSC.

Table 11.4: A summary of the key features (or 'rules of combination') of the National Senior Certificate

Compulsory subjects	
Languages	Any two official languages, one at home language (first) and one at first additional (second) language level, one must be the medium of instruction.
Life Orientation	Learners must pass a 10-credit Life Orientation course. This course is to be examined at a school level only.
Mathematics	Learners must enrol for either Mathematics or Mathematical literacy.
Optional subjects	Learners must choose any three. In order to qualify to apply for university entrance, learners must choose from a list of designated subjects and meet the minimum ratings as prescribed.

The assessment process

Nature of the assessments

The NSC has been amended a few times since the first examinations in 2008. In 2012, a Curriculum and Assessment Policy document was introduced in grade 10 for each subject. The National Curriculum and Assessment Policy Statement (CAPS) is a single, comprehensive and concise policy document, which has replaced the Subject and Learning Area Statements, Learning Programme Guidelines and Subject Assessment Guidelines for all the subjects listed in the National Curriculum Statement Grades R–12.

Examinations

The NSC examinations are conducted by three assessment bodies: the DBE; the Independent Examination Board (IEB); and the South African Comprehensive Assessment Institute (SACAI).

The DBE papers are set and moderated at the national level. The setting of examination question papers is underpinned by national standards that are embodied in the CAPS and accompanying guideline documents (DBE, 2015). One of the criteria specified in CAPS is the weightings of the cognitive demands based on the revised Bloom's taxonomy. The spread of questions in the examination paper is expected to comply with the specified weightings of the cognitive demands in CAPS. For the 2016 examinations, a total of 132 papers of the 58 subjects were set and moderated by DBE.

The administration of public examinations is a joint responsibility between the DBE and the nine Provincial Education Departments (PEDs). The DBE has a responsibility to set national standards and to coordinate and monitor the administration of the examinations across the nine PEDs. This the DBE does by the development of national policy for the conduct, administration and management of national examinations, and it also monitors the entire examination cycle from its inception to its conclusion.

The PEDs are responsible for the administration of the examination, which includes the registration of centres and candidates; the printing, packing and distribution of question papers; the writing of the examination; the marking of the examination answer scripts; and the capture of the marks on the Integrated Examination Computer System (IECS). The DBE takes final responsibility for the processing of the results, together with the standardization of the results, a process which is the mandatory responsibility of the quality assurance council, Umalusi.

The appointment of markers is governed by the Employment of Educators Act, specifically, the Personnel Administrative Measures (PAM). The PAM criteria for the appointment of markers include the following:

(a) a recognized three-year post-school qualification which must include the subject concerned at second- or third-year level or other appropriate post-matric qualifications;
(b) appropriate teaching experience, including teaching experience at the appropriate level, in the subject concerned;
(c) language competency; and
(d) in addition to the above criteria, preference is given to serving educators who are presently teaching the subject concerned.

In 2016, a total of 657,447 full-time candidates registered for the NSC examinations. The examinations were administered and marked at 6,797 and 118 centres respectively. A total of 35,000 markers were involved, for a period of two weeks, with the marking of about 10.5 million scripts.

The IEB papers are set and moderated according to the examination requirements which are detailed in the IEB Subject Assessment Guidelines and are based on CAPS. The guidelines, among other things, specify the content/topics which are examinable and the level of cognitive demands at which items should be set. For the 2016 examinations, the IEB set 81 papers in 44 subjects.

The IEB is responsible for the registration of learners, administration of exams, appointment of markers, marking of scripts and the capturing of marks. The marking of scripts is centralized and takes place in Gauteng.

In 2015, 10,212 full-time candidates from 200 schools across South Africa wrote the IEB NSC examinations. In 2016, around 11,000 full-time candidates wrote the IEB examinations. Annually, about 1.5 per cent of the Grade 12 candidates write IEB examinations.

SACAI is a new independent assessment body whose examinations papers are set and moderated according to CAPS, which is considered a minimum standard. SACAI in 2016 set and moderated internally a total of 47 papers in 27 subjects. In 2015, SACAI administered examinations to about 1,000 candidates. SACAI, like the IEB, is also responsible for the registration of learners, administration of exams, appointment of markers, marking of scripts and the capturing of marks. Assessment bodies are allowed to 'buy' papers from one another. This process is informed by low enrolments in particular subjects for a given assessment body. For instance, all papers for foreign languages are set, moderated internally and marked by IEB.

It is important to note that none of the assessment bodies use pre-standardized items for their papers. So items are not field tested before being included in the paper. Teams of examiners set the papers annually. After

the papers are set, internal moderators then approve them. The internal moderators are subject experts, different from examiners, employed by assessment bodies. One of their responsibilities is to verify or check whether the paper complies with the prescripts and standards of the exam.

Umalusi, as a quality council, externally moderates all the examination papers from DBE, IEB and SACAI. Umalusi is mandated to ensure that the NSC examinations conducted each year comply with policy. This function is carried out by the subject experts contracted by Umalusi. In addition to the moderation of papers, Umalusi also:

- monitors the writing of the examinations
- monitors and verifies marking
- verifies the capturing of marks.

School-based assessment (coursework)

The school-based assessments (SBA) are set and marked at school level. The standards for the assessments are prescribed in the CAPS for DBE and SACAI. For IEB, the SBA standards are prescribed in the IEB Subject Assessment Guidelines. Assessment bodies sometimes set common tasks as a way of supporting schools. The SBA marks are subjected to different layers of moderation. However, despite such measures, some teachers are still strict while others are too lenient with the marking of SBA tasks. As a result, Umalusi statistically moderates the final SBA marks of all learners. The moderation is relative to the learner performance in the external examinations.

Weighting of school-based assessment and examinations

All learners must comply with the internal assessment requirements associated with the NSC. Subject information required by Umalusi for determining a candidate's results must include the internal assessment mark (25 per cent) as part of the final standardized marks.

Umalusi certifies qualifications on the basis of an SBA assessment component and an external examination, both of which contribute to the candidate's final result. The SBA and external examinations are central to the quality assurance processes and are both mandatory at exit levels.

Weighting for SBA and external assessment is 25:75. Weightings for subjects with a practical assessment task (PAT) are different. The PAT allows learners to be assessed during the year and allows for the assessment of skills that cannot be assessed in the written exam (DBE, 2016).

Table 11.5 indicates the weighting for internal and external assessment in subjects with a practical assessment component.

Table 11.5: Weightings for practical assessment subjects

Assessment tasks	End-of-year assessment
SBA (School-based Assessment)	Exam Paper = 50%
	Practical Assessment Task (PAT) = 25% [If the subject has a PAT component]
25%	**75%**

Standard Setting/Standardization process

The learner's final mark consists of 75 per cent examination mark plus 25 per cent of the SBA mark. The final mark is expressed in terms of 7 levels of performance, of which 7 is the highest and 2 the lowest. Level 1 is regarded as a fail.

The examinations marks and SBA marks are standardized separately. The SBA marks are standardized relative to the standardized examination mark. This process is done programmatically after standardization or statistical moderation of the examination marks.

In principle the statistical moderation of examination marks (or standardization process) is still done exactly as it was in 1992 by Umalusi's predecessor, SAFCERT. Through the statistical adjustment (standardization) of exam marks, Umalusi aims to ensure as far as possible that the standard of the results is a reliable indicator of candidates' performances relative to those of previous years.

The standardization process is based on the principle that when the standards of examinations (from one year to the next) are equivalent, there are certain statistical mark distributions that should correspond (or be the same apart from chance statistical deviations). The assumption is also that if the nature of the cohort sitting for the examinations each year does not change, then the results should not change either.

Statistical moderation consists of comparisons between the mark distributions of the current examination and the corresponding average distributions over a number of years, to determine the extent to which they correspond. If there is good correspondence, then it can be accepted that the examinations were of equivalent standard. If there are significant differences, then attempts are made to ascertain the reasons for those differences. On occasion differences may be due to factors such as a marked change in the

composition of the group of candidates enrolled for a particular subject, poor preparations for the examinations by candidates because of some disruption in their school programmes, or, unusually, thorough preparation by candidates because of special initiatives on the part of the educators or support structures.

In the absence of strong indications that there are valid reasons for differences, it is generally accepted that the differences are due to deviations in the standards of the examination or the marking, and the marks are adjusted to compensate for these deviations.

The Assessment Standards Committee (ASC) on behalf of Umalusi Council carries out the standardization process. The committee is composed of:

- statisticians with relevant experience and knowledge in handling of statistically oriented research projects
- professionals in education with specific knowledge and expertise in assessment and curriculum.

The ASC conducts national standardization meetings for all examinations per examination cycle, per assessment body and per qualification. For the NSC and other qualifications, the meetings are in December, in most cases between 16 and 24 December. Assessment bodies are invited on the first day of the meeting to present intervention strategies implemented that might have had an impact on the performance of the learners in the current year.

The meeting is followed up by *pre-standardization* meetings that are for ASC members. At the *pre-standardization* meetings, information from two standardization booklets containing statistical information is considered. The first booklet contains:

- a historical average (norms) constructed from learners' performance in the subject. A subject might consist of more than one component. Norms are based on the raw mark distributions in the subject, averaged over the past five years. In a case where a distribution contains outliers, the historical average will be calculated excluding data from the outlying examination sitting; however, the distribution, which contains an outlier, will remain part of the three to five examination sittings
- the raw mark distributions and the cumulative frequency distributions for each of the past five years' examinations including the outlier
- the raw mark distribution and the cumulative frequency distribution of the current examination. The raw mark distributions are in terms of deciles

- the mean and median of each distribution
- Ogive graphs for the cumulative mark distributions of the current year and the previous two years.

The second booklet contains the pairs analysis and the raw mark and cumulative raw mark distributions per mark. The pairs analysis indicates performance of a cohort of learners who sat for two different subjects, such as performance of learners who sat for both mathematics and physical sciences. The performance is in terms of averages of means and medians as well as correlations.

In addition to the two booklets, chief markers and Umalusi's moderators' reports on marking are presented to the committee. The research unit of Umalusi also presents research findings from research projects relevant to the process.

The first thing that is considered before taking a particular decision is the current performance compared to the norm. In other words, the statistical information is fore-fronted. If there are no significant differences, in terms of the ogives (graph), means, medians and pass rates, then the results are accepted. If there are differences, then all the information from qualitative reports on the subject is considered. The committee members debate and persuade each other on the basis of other information available at that time. The ultimate decision is reached through consensus.

It can be concluded that the standardization process used in South Africa is a form of *cohort referencing*:

> A comparison between the mark distributions of the current examination and the corresponding average distributions of a number of past years, to determine the extent to which they correspond. If there is good correspondence, it can be accepted that the examinations were of an equivalent standard. On occasion, if there are significant differences, the reasons for those differences are established (Umalusi, 2015).

A pre-standardization meeting is followed by a standardization meeting between the ASC and representatives of the assessment bodies. In the case of the ministry, the representative is the director general, who is the highest official reporting to the minister. The independent assessment bodies are represented by their chief executive officer. The meetings are chaired by the chairperson of Umalusi Council and are also attended by other stakeholders such as teacher unions. At the meetings, the assessment bodies present their adjustment proposals per subject. If a proposal for a subject is the same as

the one taken by the committee at its pre-standardization meeting, then it is accepted. If the proposal is different from but not better than the committee's decision in terms of pass or failure rates, then the ASC decision is usually accepted without reservations. While the ASC can provide a rationale for its decision and can also be persuaded to consider other information at the meeting with the assessment bodies, its decision is considered final.

Due to a tight standardization schedule, the standardization process cannot accommodate appeals. In other words, assessment bodies cannot appeal the standardization decisions. However, all assessment bodies have re-marking processes in place for learners to appeal if not satisfied with examination mark.

Political and public controversies/debates

The NSC is a gateway qualification that allows learners access to higher education institutions and the world of work. As such, the annual announcement of the Grade 12 learners pass rates in South Africa is always received with scepticism from universities, political commentators and the public. This is simply because of the purposes of the NSC.

For years, South African universities accepted the matriculation certificate (Senior Certificate – SC) as the best single predictor of academic success at tertiary educational institutions. Scott *et al*. (2007) indicated that universities relied on the SC exams for admission purposes because of their proven track record as a relatively robust signal of student success at institutions of higher learning. However, the introduction of a new qualification, the NSC, which was examined for the first time at the end of 2008, created uncertainties among universities. In particular, higher education institutions questioned whether the NSC was actually an improvement on the former SC, as well as the NSC's ability to predict academic success at tertiary level. Nel and Kistner (2009) stated that a major concern with the introduction of the NSC was the stipulation of the standard for examination question papers in 2008 in light of the scrapping of grade levels. Unlike with the SC, where learners could be examined at two different grade levels (higher or standard grade), the new qualification, the NSC, is only examined at one level. There is no longer a distinction between subjects on a higher or standard grade level.

This response arose mainly because in 2008 the NSC exam produced an unusually high number of students who qualified for university admission. As a result, in 2009, universities experienced an abnormal influx of first-year students, and several institutions complained of higher-than-normal pass rates.

Several studies in various university disciplines and at various tertiary institutions in South Africa have been carried out to determine the preparedness of students who wrote the NSC, as well as their subsequent success rates. Most of these studies illustrate that NSC results have a lesser ability to untangle academic performance at university level, shown by weak correlations between NSC results and university performance. Recent research has shown that NSC scores are inflated by about 20 per cent and are thus poor predictors of first year achievement in Economics at the universities of the Witwatersrand (Schöer *et al*., 2010; Hunt *et al*., 2011), Stellenbosch (Nel and Kistner, 2009) and the Western Cape (Dlomo *et al*., 2010).

However, most of these studies concentrate on a particular year, specific courses and programmes of specific universities, and differ in the choice of the dependent variable. Therefore, these studies tend to be limited to a very specific sample of students and do not provide a picture across time and across different institutions that can illustrate the ability and the trend in the ability of NSC matriculation marks to act as predictors of academic success at higher education institutions in general.

The scepticism from universities and political commentators is unfortunately directed in the main to the public system. The scepticism is also fuelled by the implementation of the policy on progression and Umalusi's position on language compensation.

The implementation of the Progression Policy in Grades 10–12 has attracted considerable attention from various quarters. The policy on progression, while it has been implemented in the lower phase for years, was only recently enforced in the further education training (FET) phase. The FET phase is Grade 10 to Grade 12. In terms of the policy, a learner may only be retained once in the Further Education and Training Phase in order to prevent the learner from being retained in this phase for longer than four years. Policy on progression has been applied in the FET band since 2013. But this policy has been applicable in the general education and training band since Curriculum 2005. The rationale for the policy is that:

> South Africa loses half of every cohort that enters the school system by the end of the 12-year schooling period, wasting significant human potential and harming the life-chances of many young people. Secondary school completion rates are at 77% in the United States, 87% (to the age of 16) in the United Kingdom and 93% in Japan. South Africa should aim for a comparable completion rate of between 80–90% (Poliah, 2016).

On language compensation, in 1998 a team of researchers, appointed by the then Minister of Education, concluded that learners who write Senior Certificate Examination (SCE) in a language that is not their mother tongue are seriously disadvantaged. Note that the examinations are in two languages, English and Afrikaans. However, the majority of learners, 80–85 per cent, of those who write the examinations have English or Afrikaans as a second language. The researchers further proved that language was or is a major factor contributing to poor performances by such learners.

SAFCERT (now Umalusi) decided in 1999, as part of its responsibility to ensure fairness in the SCE, to apply a *compensatory measure* for learners whose first language was neither English nor Afrikaans and who offered an African language as their first language. A compensation of 5 per cent was awarded to such learners for the non-language subjects, based on the mark they had obtained.

According to Umalusi (2004), the compensatory mechanism was implemented as an interim measure while the provincial departments were in the process of upgrading the teaching and learning of English Second Language. It was agreed in principle that as the proficiency levels in English Second Language improve, this compensatory measure will be reviewed.

In 2012 Umalusi conducted further research on the language compensation practice as part of the NSC. Based on the findings, it was decided to gradually decrease the compensation rate by 1 per cent yearly to 0 per cent in 2018. However, the decision was again reviewed in 2016 and it was agreed that it be fixed at 3 per cent for now.

References

DBE (Department of Basic Education) (2016a) *Evidence Based Report*. Internal report. Pretoria.

DBE (Department of Basic Education) (2016b) *National Senior Certificate Report*. Internal report. Pretoria.

Dlomo, Z., Jansen, A., Moses, M. and Yu, D. (2010) 'Is performance under the new matric curriculum still significant in predicting first year academic success in economics?'. Paper presented at Indaba Hotel and Conference Centre. 27–29 October. ESSA Conference. Online. https://goo.gl/iyWZ8g (accessed 18 June 2018).

Hunt, K., Ntuli, M., Rankin, N., Schöer, V. and Sebastiao, C. (2011) 'Comparability of NSC mathematics scores and former SC mathematics scores: How consistent is the signal across time?'. *Education as Change*, 15 (1), 3–16.

NECT (National Education Collaboration Trust) (2015) *Perspectives and Lessons on Public System Improvement: The case of the national examinations system*. Pretoria. Online. https://goo.gl/AEA82a (accessed 18 June 2018).

National Planning Commission (2011) *National Development Plan: Vision for 2030*. Pretoria: Department of the Presidency. Online. https://goo.gl/ZMhnqu (accessed 18 June 2018).

Poliah, R. (2016) *Progress learners and multiple examination opportunity*. Presentation to the Executive Committee of Umalusi. Pretoria: Department of Basic Education. Online. https://www.ru.ac.za/media/rhodesuniversity/content/institutionalplanning/documents/NPC%20National%20Development%20Plan%20Vision%202011.pdf (accessed 15 July 2018).

Schöer, V., Ntuli, M., Rankin, N., Sebastiao, C. and Hunt, K. (2010) 'A blurred signal? The usefulness of National Senior Certificate (NSC) mathematics marks as predictors of academic performance at university level'. *Perspectives in Education*, 28 (2), 9–18.

Scott, I., Yeld, N. and Hendry, J. (2007) *A Case for Improving Teaching and Learning in South African Higher Education*. Higher Education Monitor No. 6. Pretoria: Council on Higher Education. Online. https://goo.gl/cBxujn (accessed 18 June 2018).

Statistics South Africa (2016) *Community Survey* (Statistical Release P0301). Pretoria: Statistics South Africa. Online. www.statssa.gov.za (accessed 1 May 2018).

Terblanche, J.D.V. (1989) 'Official developments in the field of education since the De Lange report'. *Orientation*, 3 (55/57): 44–55.

Trümpelmann, M.H. (1991) *The Joint Matriculation Board: Seventy-five years achievements in perspective*. Cape Town: National Book Printers.

Umalusi (2004) *Investigation into the Standard of the Senior Certificate Examination: A report on the research conducted by Umalusi*. Pretoria: Umalusi Council for Quality Assurance in General and Further Education and Training. Online. https://goo.gl/KaYUZP (accessed 18 June 2018).

Ambitious objectives and persistent challenges: National examinations in post-apartheid South Africa

Sarah Howie

This chapter is a valuable contribution to the literature in describing and demystifying the standards setting process in relation to the National Senior Certificate (NSC) in South Africa. It presents a good description of the South African national examination system related to the end of secondary school national examinations, the NSC, providing an interesting model of standard setting in a complex emerging context. The NSC is the highest stakes examination in the country and causes a number of unintended consequences (Howie, 2012). The chapter describes the landscape of the South African education system broadly and of the examination system in greater depth, including a few debates raised nationally about the examination. The history provided in this chapter is essential to understanding the developments in the system over the past 100 years and more. What is implied but not as clear in the historical description is the severe impact of the apartheid system on the examinations and the differentiation in quality as a result (Howie, 2003, 2015). Previously, different racial groups attended separate and different institutions managed by 19 diverse education bodies and thus wrote different examinations with considerably varying standards. This is critical to understanding the challenges existing in the current standard setting and examination system in general today. The chapter then describes the assessment process conducted in the NSC. Presumably the NSC was selected as the case study given its position as the largest of the examinations conducted in South Africa and because of its high stakes nature.

The first common set of examinations set and administered nationally is a very recent event (since 2008) compared to most countries, and therefore many teething problems were inevitable as the national system found its feet. The first phase of the centralization process from 2008 was characterized by perceived low standards and irregularities such as large-scale examination paper leakages with the complicity of department staff. The leakages have

reduced significantly in recent years, with more localized irregularities emerging such as isolated cases of group-copying with teacher involvement (DBE, 2014). Tougher measures and criminal charges being implemented have assisted in reducing, although not eradicating, this behaviour. Considerable attempts have been made to improve the standard of the NSC papers. However, the need for capacity development after apartheid is still substantial as the system is hampered by the lack of capacity from the classroom teaching to the systemic level regarding setting and moderating papers, to marking examination papers, to the quality assurance of the entire process (DBE, 2014), the impact of the differentiated systems under apartheid still haunts all levels of the education system.

Another development at the national level was the moving away from the reliance solely on the statistical intervention during the standard setting process. The quality assurance body merged two separate committees. The previous Statistics Committee traditionally dealt with the standardization of the results and included experienced experts, mostly statisticians, in the process. The Assessment Committee comprised practitioners with expertise in assessment from universities in education, including adult and vocational education. The merging of these two committees was beneficial is some ways, but the unintended consequence was that the standardization process lost expertise and emphasis on the statistical standardization in the process. The emphasis shifted to a consensus model and capacity development. While the volume of qualitative data was dramatically increased to the benefit of the process in general, much of this data is difficult for members to digest in a very short period during the standardization process.

The chapter noted the removal of higher and standard grade differentiation in the transition to the NSC in 2008. This inevitably led to the production of easier papers and their inability to discriminate sufficiently within one paper (Howie, 2016). Despite warnings about the consequences (DBE, 2014), this has resulted, for example in one school, where one third of learners obtained 90 per cent aggregate and 60 per cent obtained 80 per cent aggregate with many questioning the standard of the papers as mentioned in the chapter. While there are many national commentators and armchair experts, insufficient research has been conducted on the NSC as indicated in the chapter and reflected in the limited national references and a dependency on Umalusi research.

Another important factor raised by the chapter affecting the examinations is the language of instruction, which has a significant impact on the quality of education in general (Howie *et al.*, 2017) and on examinations in particular (DBE, 2014). Education is offered in all 11

official languages from Grades 1 to 3, and thereafter from Grade 4 only in Afrikaans and English. Given that more than 80 per cent of the learners do not speak these languages at home, this has been found to have a significant effect on learner performance in the NSC (DBE, 2014). While a language compensation measure was originally introduced as a temporary measure, it has been retained in the system despite recommendations for its removal. Although calls have been made repeatedly for improved language development strategies, the system to date has not implemented a systemic remediation intervention nor succeeded in improving the language proficiency of the teachers or learners.

The chapter does not raise or problematize the issue of the lack of capacity in South Africa regarding assessment and examinations in particular and affecting standard setting. There is a dire shortage of suitably qualified and trained personnel in psychometrics and assessment in education as well as few professional statisticians working in and understanding education. This shortage has had a negative impact on the country, putting strain on the ability of the examination bodies as well as the quality assurance institutions. This lack of capacity results in political rather than expert judgements influencing outcomes at times within the system (Howie, 2016).

Not mentioned in this chapter is the Ministerial Committee tasked with reviewing the quality of the NSC (DBE, 2014), which revealed a number of shortcomings with the current examination systems. While acknowledging the progress in the national system given its ten years of existence, nonetheless the current challenges regarding the quality of the examinations and of the quality assurance were noted. Hints of these are found in this chapter. Key to addressing these is developing competence of the actors involved from the examination panels, moderators of the papers, the personnel and committees within the examination bodies and quality assurance body. There is still a significant amount of work to be done to achieve a more valid and reliable standard setting system, but the system has come a long way. This chapter is an important contribution towards explaining the processes behind the NSC and therefore towards the goal of achieving an effective standard setting system in South Africa.

References

DBE (Department of Basic Education) (2014) *Ministerial Task Team Report on the National Senior Certificate (NSC)*. Pretoria: Department of Basic Education. Online. https://goo.gl/ZXsSpD (accessed 18 June 2018).

Howie, S.J. (2003) 'Language and other background factors affecting secondary pupils' performance in mathematics in South Africa'. *African Journal of Research in Mathematics, Science and Technology Education*, 7 (1), 1–20.

Howie, S. (2012) 'High-stakes testing in South Africa: Friend or foe?'. *Assessment in Education: Principles, policy & practice*, 19 (1), 81–98.

Howie, S.J. (2015) 'What do the IEA studies mean for developing countries education systems and educational research?'. Keynote address presented at the 6th IEA International Research Conference, Cape Town, 24–26 June. Online presentation. https://goo.gl/1EuyaA (accessed 18 June 2018).

Howie, S.J. (2016) 'Assessment for political accountability or towards educational quality?'. Keynote address presented at the International Association for Educational Assessment (IAEA) Conference, Cape Town South Africa, August.

Howie, S.J., Combrinck, C., Roux, K., Tshele, M., Mokoena, G. and Macleod Palane, N. (2017) *Progress in International Reading Literacy Study 2016: South African children's reading literacy achievement*. Pretoria: Centre for Evaluation and Assessment, University of Pretoria. Online. https://goo.gl/xHJjLT (accessed 18 June 2018).

Improving standards or establishing (or developing) performativity regimes?

Anil Kanjee

In Chapter 11, Sibanda provides a brief history of examinations in South Africa, illustrating how the political transformation process has impacted on the examination system. The chapter highlights key issues regarding different examinations bodies, certification requirements and reporting specifications, admissions into higher education and the standardization and the compensatory measures applied to examinations results. This commentary focuses on the use of the matriculation examination results as 'the standard of education' and its impact on learning and teaching in schools.

The Grade 12 examinations, popularly referred to as the matric exams, are extremely high stakes national examinations taken by all learners upon completion of schooling. Since its primary purpose is to certify learners' competency to enter the labour market and/or the higher education sector, success or failure in this single examination has a significant impact on the life trajectory of all South African children (Reddy, 2006). In this respect, the matric exam has maintained its key function, despite the significant changes that have impacted the country and the education sector over the last century (Kanjee, 2006). However, the characteristics of the examinations process, as well as its impact on the education system, have changed dramatically over the years.

That the matric examination results are viewed as a measure of 'the standard of education' in the country is not surprising given the globalization of performativity and accountability regimes, and the absence of any performance measures at the secondary education level in South Africa (Chisholm and Wildeman, 2013). This has resulted in holding schools and districts accountable for learner performance and has manifested in several ways. First, the release of the results has focused specifically on year-on-year improvements in pass rates, promoting an annual horse race among provinces to be 'Number 1'. Second, the results and names of schools and districts with low pass rates have been made public, increasing the pressure to produce higher pass rates. Third, schools and districts deemed as

performing below 'the standard' have been targeted for specific intervention as well as greater monitoring by provinces and districts.

Notwithstanding the increased focus on and investments in secondary schools, the impact of this approach has largely been detrimental, with learners from poor and marginalized backgrounds bearing the brunt. To increase their pass rates and meet minimum thresholds to avoid being classified as dysfunctional, most schools have focused specifically on improving pass rates by teaching to the examinations, rather than on improving learning and teaching. More concerning, many schools have also resorted to retaining Grade 11 learners that they believe may not succeed in Grade 12, thus creating additional challenges in Grade 11, while also encouraging learners to select 'softer subject options', or to enrol as private candidates (Chisholm and Wildeman, 2013; Motala *et al.*, 2009).

District officials have also instituted several measures for improving pass rates that include providing additional classes to learners, usually after school, and/or during weekends or holidays. Often, these classes end up as drill sessions that teach to the expected content of the exams. In addition, it is common practice for districts to prioritize the matric examinations during the school year as well as to allocate all subject advisors, even those responsible for primary schools, to monitor the matric examinations, effectively limiting support provided to many schools while also terminating support during the examinations period (Mavuso, 2013).

Universities have also questioned the use of the matric results as a valid measure for admissions and for predicting success within the higher education sector. Specifically, universities have argued that most learners passing the matric examinations are under-prepared to enter the sector, resulting in high percentages of students dropping out or failing to complete their degrees (van Broekhuizen *et al.*, 2017). In responding to this challenge, universities have implemented the National Benchmark Tests, which are used to identify students in need of additional support and as an alternative admissions process (le Roux and Sebolai, 2017).

While the matric examinations play a valuable role in South African society, their use as 'the standard' against which to hold schools accountable has had a detrimental impact on the education system, with learners from poor and marginalized backgrounds bearing the brunt of its negative impact.

References

Chisholm, L. and Wildeman, R. (2013) 'The politics of testing in South Africa'. *Journal of Curriculum Studies*, 45 (1), 89–100.

Kanjee, A. (2006) 'Comparing and standardising performance trends in the matric examinations using a matrix sampling design'. In Reddy, V. (ed.) *Marking Matric: Colloquium proceedings*. Pretoria: HSRC Press, 72–89.

le Roux, N. and Sebolai, K. (2017) 'The National Benchmark Test of quantitative literacy: Does it complement the Grade 12 Mathematical Literacy examination?'. *South African Journal of Education*, 37 (1), 1–11.

Mavuso, M.P. (2013) 'Education district office support for teaching and learning in schools: The case of two districts in the Eastern Cape'. Doctoral thesis, University of Fort Hare. Online. https://goo.gl/RmDzNF (accessed 18 June 2018).

Motala, S., Dieltiens, V. and Sayed, Y. (2009) 'Physical access to schooling in South Africa: Mapping dropout, repetition and age-grade progression in two districts'. *Comparative Education*, 45 (2), 251–63.

Reddy, V. (2006) 'Introduction'. In Reddy, V. (ed.) *Marking Matric: Colloquium proceedings*. Pretoria: HSRC Press, xii–xix.

van Broekhuizen, H., van der Berg, S. and Hofmeyr, H. (2017) *Higher Education Access and Outcomes for the 2008 National Matric Cohort*. Stellenbosch Economic Working Papers No.16/2016. Online. https://ssrn.com/abstract=2973723 (accessed 18 June 2018).

Chapter 12

Standard setting in Sweden: School grades and national tests

Christina Wikström and Anna Lind Pantzare

Standards in a Swedish educational context

The term 'standard' generally refers to a certain quality or performance level, or something commonly agreed. According to the Swedish Standards Institute, a standard is 'a document, set up/prepared in consensus with and by an acknowledged institution or organization, that for public and repeated use will define rules, guidelines or characteristics for activities or their outcomes, with the purpose of achieving order so far as possible in a certain context' (Swedish Standards Institute, 2016, authors' translation).

In a Swedish educational context this translates naturally to the National Curriculum, which is issued by the National Agency for Education (NAE), on behalf of the Swedish government. The National Curriculum is complemented with separate documents such as syllabi and grading criteria for subjects and courses. Together with the Swedish Education Act, the National Curriculum and its attachments regulate all Swedish schools, from pre-school to upper secondary school. The documents state the schools' mission, values and goals, and give directives in terms of what the schools are to do, what to teach and what to assess. Consequently, since the Swedish system is grounded in these documents, it can be described as standards-based. However, in education contexts the term 'standard' has various definitions and sometimes highly debatable meanings, which often complicates discussions about education and assessment. Standards can have to do with performance levels in grading criteria and for tests, but also refer to outcomes – that is, student and school performances, and to what extent assessment and grading can be seen as valid and reliable performance measures, within and between schools, and over time. When it comes to maintaining standards, the Swedish standards-based system is less straight-forward; research and evaluations have shown that there are fluctuations especially when it comes to how grading criteria are interpreted

and methods for assessing what the students know and can do (see, for instance, Klapp Lekholm, 2008; Gustafsson and Erickson, 2013; Tholin, 2006; Skolinspektionen, 2011; Vallberg Roth *et al.*, 2016).

This chapter will describe the Swedish criterion-referenced and standards-based system, with special attention to how assessment and grading is carried out, with a section focusing on the national tests as important elements for reliable and valid grading. We will also discuss problematic issues related to monitoring and maintaining outcome standards.

Sweden and Swedish education

Sweden is one of the Scandinavian countries, and a member of the European Union. The population is currently 10 million, and, with the exception of a handful of larger cities, the country is relatively sparsely populated. Economically and socially, the Swedish system follows the Nordic model, with a combination of free market capitalism and a comprehensive welfare state. There is a high general taxation, but also a high degree of social tax returns and public services in the form of free health care, an extensive social-service system and free education.

Sweden has a history of having a centrally regulated and coherent school system. Although it has changed in many ways over time, some fundamental elements have remained. Typical for the Swedish system is a strong belief in free education, equal opportunities and lifelong learning, and typical for educational policy is an ambition to combine regulation with freedom. While there is a general belief in the necessity of having central guidelines and standards, there is also a belief in local responsibility, giving the schools freedom when it comes to methods for teaching and assessment. This is particularly the case when it comes to assessment and grading: the teachers have the sole responsibility for assessing and grading their students, and are to base their grading on observations and other evidence collected in the classroom. Another characteristic that is especially relevant in this context, and for the discussion in this chapter, is that there has been, and still is, a general resistance to grading, standardized high stakes tests and external examinations in education, especially when it comes to younger students.

Brief outline of schooling system

The Swedish school system is structured as follows: all schools are regulated by the government and government agencies. The Ministry of Education decides on laws and educational targets, and the NAE is responsible for carrying this out in practice and to make sure that the schools are informed

of what they should teach and what the regulations are. However, although the system is centrally regulated, the schools on elementary and upper secondary level are run by the municipalities or private vendors, who run the so-called 'free schools'. All schools, including the free schools, are tax-payer funded and financed through a voucher system. Fees are not allowed. Since 2009, there is also a Schools Inspectorate, which monitors that the schools' work is in line with the regulations.

The educational system comprises non-compulsory pre-school (until the age of six) followed by nine years of compulsory education. Students normally graduate from compulsory school at the age of 16. Thereafter, most students continue to three years of upper secondary education, where there is a wide variety of programme orientations that can be divided into programmes with a vocational focus, and programmes for students on an academic track. All programmes are expected to give basic eligibility to higher education, although in vocational programmes this has to be done through additional course choices. Most are so-called national programmes that, in theory, are to be comparable in format and content across the country. Still, although there is a basis for comparability, there are also differences and variations. The programmes are not strictly standardized, and there is some degree of freedom for the schools to decide on. All programmes include a fairly large number of subjects and courses. Core subjects, such as Swedish, English, mathematics, social science, history and natural science, are compulsory, while the weight of these subjects (the presence of advanced courses) and additional subjects depend on the programme. There are also other local variations that fall outside the regulations, such as school profile, classroom didactics or teacher quality.

Assessment and grading

As previously mentioned, the Swedish system is characterized by a combination of strong regulation and local authority. This is perhaps especially prominent when it comes to assessment and grading. There is also a tension between a belief in the usefulness of statistics and educational measurement on the one hand, and on the other, a resistance towards testing and the 'labelling' of students. There are historical and cultural reasons behind this, and the two paths can be visible also in the current educational system. From a historical perspective, the views on and methods for assessing students' knowledge and skills have varied, and to a large extent also reflect current ideological trends in society.

At the beginning of the comprehensive school system, assessment and grading was for the most part a local concern. However, in the post-war

expansion of secondary and post-secondary education, there was a need to find fair and reliable instruments for credential purposes and for the selection to further education. The mid-1900s was an era characterized by a strong belief in measurement and statistics, and both scholars and policymakers were influenced by psychometric research, especially from large-scale testing in the US. The idea of a 'cohort-referenced' grading scale, based on a normal distribution, was suggested as the solution. The idea was to make grades comparable, and also to make it possible to calculate a grade point average (GPA) that could be used for ranking the students when applying to higher education. The cohort-referenced grading system was adapted throughout the school system during the 1960s, with a scale ranging from 1 to 5 where 3 represented average performance.

The main idea with the cohort-referenced scale was easy to understand, but many teachers found it complicated in practice. For instance, a common misunderstanding was that the scale was to be based on the relative positions in the classroom, which made it more difficult to get a high grade in a high performing class, and vice versa.

The educational reforms that were the result of a long-term ambition to introduce standardization and reliable measures of performance clashed with a new era of radical movements and criticism towards the established system and traditional forms of education and assessment. When the cohort-referenced system was introduced in upper-secondary level, end-of-school exams were abolished and the responsibility of grading the students was given to the teachers. Standardized tests were made available to the teachers to provide information on their students' positions on the scale (the cohort distribution). Apart from the inconsistencies in grading, the grading system itself was criticized from several perspectives. Many viewed grades as negative for the students and their learning, and the cohort-referenced grades were found particularly problematic, as students (and teachers) often were more focused on how they performed relative to other students than on what they actually learnt. From a policy perspective, the cohort-referenced grading system was found lacking since it made educational evaluation difficult, especially when wanting to make comparisons over time, as the whole idea was that the mean and deviations would always be the same. This system is described in more detail in Andersson (1991) for instance.

It was argued that the grading should be abolished altogether, at least for elementary school, or that a goal or criterion referenced system should be introduced instead (Wedman, 1983: 2000). From 1969 onwards, students were graded less frequently than before, and a trial period was introduced (due to political controversies) that made grading in primary and lower

secondary school non-compulsory. About 50 per cent of the municipalities decided to abolish grading until 8th grade, when the students were 15 years old. In 1980, it was decided that no students should be graded until the end of Year 8, and this remained until 2014, when grading at the end of lower elementary school (Year 6) was introduced.

Controversial to the general assessment trends in society, a standardized test for college admission, strongly influenced by the aptitude testing in the US, was introduced in 1977. The Swedish Scholastic Aptitude Test (SweSAT) differed from its American model, however, since it was only open for older students (25+) with work experience, and introduced as a way to broaden the recruitment to higher education. The SweSAT and its background are described in detail by Wedman (2017). In 1991 the test became open for all, to function as a 'second chance' to those whose grades from upper secondary school were not high enough. When there is a selection among eligible applicants, the universities must admit at least 30 per cent of the applicants from this group.

The late 1980s and early 1990s was a period of political turmoil, with strong neo-liberal trends in society. There was a growing belief in privatization, competition and deregulation, and after a change in government an extensive reform programme began (see, for instance, Blanchenay *et al*., 2014). In only a few years' time, the Swedish school system went from being one of the most centralized and regulated systems within the OECD to one of the most decentralized and deregulated (Lundahl, 2002). The responsibility for running the schools was moved from the state to the municipalities, and the so-called 'free school reform' opened up for private, or independent, schools. A voucher system was introduced, making schools compete for their students, for example in regard to funding (Parding, 2011).

In 1994, the criterion-referenced system that had been discussed for decades was finally implemented. A new national curriculum was introduced, with a new grade system based on a criterion-referenced scale: IG (fail), G (pass), VG (pass with distinction) and MVG (pass with special distinction). The new national curriculum had been changed in a number of ways as the former curriculum had presented rather detailed descriptions of what should be taught in each subject, leaving limited freedom to the teachers, but with less instruction in terms of what to assess and grade. The new curriculum focused on defining goals rather than detailed content – goals to strive for and goals to achieve. The national curriculum has since then comprised three main parts: first, a document stating the common values and mission for all schools; second, the overall objectives and directives; and third, the syllabi with performance levels for each grade

level (standards). When introduced, these descriptions proved to be rather vague, resulting in severe problems for the teachers. Grading was seen as especially difficult due to unclear grade descriptors and wide grade levels, allowing large variations within the boundaries of each grade. In 2011, the National Curriculum was revised to make it clearer. The grading scale was also increased with more grade levels – the former three pass levels became five: E–A, and F (fail). See also Erickson (2017) for a description of the current system.

Qualifications needed for higher education

There are two types of eligibility for access to higher education in Sweden: basic and specific. Basic eligibility is acquired by having graduated from upper secondary education (or equivalent) and passing the courses (Grade level E minimum). Requirements for specific eligibility then depend on the chosen university programme, and often entail more advanced courses in certain subjects that are seen as specifically relevant for the programme.

When competing for study places in selective university programmes, applicants who have the required eligibility are then rank ordered based on their added grades, which is a type of weighted GPA. In the present version of the system, the criterion-referenced letter grades are transformed to a numerical (but still ordinal) scale from 0 to 20 (10 for E, 12.5 for D, 15 for C, 17.5 for B and 20.0 for A). The grade values are then calculated by course length (even though ordinal data are not strictly suitable for this), and then averaged to a GPA ranging from 0 to 20. Currently, extra merits are given for advanced courses in mathematics and second languages (other than Swedish), that can add 2.5 and make the maximum GPA 22.5.

If an applicant to higher education believes that his or her true ability is higher than what is reflected in the GPA or just wants to maximize his or her chances, there is also the option to take the SweSAT. Applicants who have taken the test are placed in two selection groups in the admissions process, the GPA group and the SweSAT group, and selected on the basis of which instrument ranks them the highest. An applicant has therefore nothing to lose by taking the test. The majority of those taking the SweSAT are around 20–21 years of age, but it is not uncommon that students in upper secondary school, especially those who aim for highly selective study places, take the SweSAT at some point before they graduate. The popularity of the test varies somewhat between administrations and over time, but tends to increase in times when there are elements of uncertainty in the grading system, and the competition for the study places in higher education is strong.

Recent reforms

Educational reforms generally take place when there is a change in government, and Sweden has changed government a few times in the last few decades. Typically, each major change in government has resulted in new national curriculums and new or revised assessment systems. The revisions have varied in magnitude and focus. Since Sweden has had a long period of socio-democratic rule, which has shaped the education system over a long period of time, the main changes have taken place during periods of right-wing or liberal governance. During these periods, there has been a trend towards more national tests, more monitoring through national tests, and earlier grading. For example, in 2009, national tests were introduced for Year 3 and Year 6, and grading introduced in Year 6. Earlier grading was also discussed, but not implemented. In 2011, revised national curricula were published, and the grading scale increased from three pass levels to the current five (E–A). This trend has not been too controversial, however, since assessment and monitoring of school performance and outcomes is generally believed to be of importance to shape up and improve a school system that seems to be decreasing in quality, based on findings from international studies such as the Programme for International Student Assessment (PISA) and the Third International Mathematics and Science Study (TIMSS: OECD, 2015). In 2011, the School Law was revised, for instance with a new regulation stating that only certified teachers should be given full rights as teachers when it comes to employment, salary and responsibilities – including grading students (SFS, 2010). This applied also to teachers already in the profession, meaning that teachers who have entered the profession from another path than the traditional teacher education have to take supplementary courses and/or professional practice. The intentions behind the accreditation were to raise educational quality and the status of the teacher profession. Teachers who, for some reason, have entered the teaching profession via other paths, for instance by being an expert in the particular subject (i.e. a chemist working as a chemistry teacher, a native speaker teaching his or her native language, a musician teaching music and so forth), have to undertake supplementary education in courses relevant for the teaching profession to receive their accreditation. There have been mixed reactions to this reform, as it has caused practical problems for some schools and the teachers affected, but it has also dramatically increased the number of university courses in teaching and assessment and opportunities for professional development for teachers. Another criticism is that this regulation has meant that accredited teachers

with non-accredited colleagues have to grade students they do not know, in subjects they do not teach.

Under the current socio-democratic regime, the focus on more assessment and grading is less prominent. Recently, some sort of compromise regarding early grading has been made, where schools are able to choose if they want to introduce grading already in Year 4. It is presented as a trial period, and so far few schools have expressed an interest in participating.

When it comes to later school years, and the transition to higher education, recent reforms have mainly had the purpose to make improvements to the components in the system, while the model for assessment and selection to higher education has remained unchanged for the most part. It should, however, be noted that recent evaluations by two commissions, the school commission and the commission for entrance to higher education, have proposed a number of changes. One of these changes is to return to the former model where students in upper secondary school are graded after each semester rather than after each course, and to calculate the GPA on the basis of end-of-school grades, rather than aggregating over time, to reduce stress and pressure for both teachers and students. The current model is criticized for not encouraging improvement since students who do not perform their highest from the start, or receive lower course grades than expected, will not be motivated to try harder – their GPA will always be affected by the lower grades. This is especially problematic for students on an academic track – if aiming for a highly selective study programme in higher education, a very high GPA is needed, and every course grade will count in this competition.

The assessment process

As previously mentioned, Swedish school teachers have the sole responsibility for assessing and grading their students. It is a regulated process, but with a lot of freedom when it comes to methodology. The National Curriculum, syllabi and performance descriptors are of key importance for making this system work, and there are also guidelines issued by the NAE. Also, in some subjects and courses, there are so-called national tests, with the main purpose to support reliable and valid grading. These tests are described in more detail below.

Having a criterion-referenced grade system is probably seen as something positive by most: it is the performance of each student that will decide the grade, not the relative position to other students. Giving the responsibility of assessment and grading to the teachers is also seen as natural, as formal end-of-school examinations are seen as something of

the past. However, most teachers probably also agree that it is a complex task, and that the criterion-referenced system has made the assessment task even more difficult. While teachers are generally good at rank ordering, and at centring allocated grades around an average grade level, it is more complicated to assess knowledge and skills on a more detailed level, and linking this to performance descriptors that are not always entirely clear, in a reliable and valid way.

Typical assessment formats

Teachers grade their students by collecting and evaluating classroom evidence. How this is done varies between teachers but also with the nature of the subject and course, and traditions within that subject. Many teachers have not been trained in assessment and grading during their teacher education, which may seem strange in a system where teachers have the authority to form these decisions, but there are cultural and historical reasons behind this. When it comes to assessment formats they are likely to be influenced by colleagues and their own experiences when choosing procedures and methods, and perhaps also by the format and content of national tests. It is common that teachers use tests they have developed themselves, and other types of written exams, such as reports and essays, but also observations made in the classroom, where teamwork and collaboration projects are not uncommon. The written tests are generally paper-based, but this is likely to change with the infrastructure available in the classrooms – most students on upper secondary level have access to laptops, and schools tend to communicate online with the students, and their parents, and rely on different kinds of eLearning support.

Determining grades

Teachers use different strategies when determining grades. The approach is generally to gather as much information as possible through a portfolio approach, where coursework, teacher observations (notes) and test scores are collected and weighed into a composite grade. It is up to the teachers to weigh each of the elements, but there is still a rather complicated instruction to the teachers in how to interpret grade criteria and what to do in case the outcome is not homogeneous. The current grading scale ranges from F (below pass – fail), and the pass scores E, D, C, B and A (highest). There are performance descriptors/knowledge requirements for grades E, C and A. Intermediate grades D and B are given when all objectives are met for the lower grade, but there are parts missing for the higher grade.

This model is often discussed, since teachers tend to interpret such instructions very differently (see, for instance, Vallberg Roth *et al*., 2016). In subjects and courses where there are national tests, the test scores are generally given a lot of weight. The grading process can be seen as more difficult for teachers who do not have the support of national tests, but they also experience less external control. The test scores are nowadays collected by the NAE, and it is often considered problematic if teachers or schools divert too much from the test scores in their grading. This also increases the stakes of the tests for the students.

The variation in assessment methods and strategies for determining grades has resulted in reliability and validity problems, which have been illuminated in research and also strongly criticized (Klapp Lekholm, 2008; NAE, 2016; Wikström, 2005), and this has increased the focus on national tests, as a way to promote fair and valid grading. See Erickson (2017) for a more thorough description of the relation between grades and national tests.

The national tests

The national tests have a significant role in the assessment process today and are available, and mandatory, for core subjects such as Swedish, English and mathematics for elementary (some years) and upper secondary school (some courses). There are also national tests in social science subjects and natural science subjects in Year 9 (15-year-olds). The format and length of the tests varies and depends on subject and level.

These tests should not be seen as traditional high stakes examinations, but they do play an important part in the grading process, where their importance, purposes and stakes for the students have been increased over time, as mentioned earlier. The tests are now also expected to provide information about goal achievement at the school level, municipal level and national level. Still, the main purpose has always been, and still is, to support comparable and fair assessment and grading. The tests are also expected to have a positive effect on teaching and learning, by making curricula and criteria for grading more concrete for both teachers and students, and, as a consequence, increase students' goal achievement. The ambition is that the tests should be exemplary and not only assess the parts that are easy to assess. This has resulted in, for example, oral parts in the mathematics tests and laboratory parts in the natural science tests.

The main part of the test administration is a local responsibility. The tests are made available to the schools and their teachers to be administered on a certain date each year. The administration is a fairly standardized procedure, with exceptions for students with special needs, who can

be allowed extra time, or a separate room. The tests are marked by the teachers; normally the teachers mark only their own students' booklets. The test scores are then aggregated and reported on the same grading scale as the grades (see description above). This has for a long time been a local affair, with no systematic moderation or external control other than random checks by the Schools Inspectorate, which sometimes has been criticized, since it has been claimed that similar bias as can be found in teachers' grading is also mirrored in the scoring of the tests (Gustafsson and Erickson, 2013; Skolinspektionen, 2011). In addition, mainly as a consequence of the criticism regarding the lack of comparability in the grading, the NAE has emphasized the importance of collaboration in the marking procedure and this seems to have become more common (NAE, 2014).

Although the national tests are owned by NAE, the tests are developed externally – with some interaction throughout the process between test developers and the NAE, who have the final say before the tests are administered and used. The task of developing the test is given to Swedish universities, usually to departments focusing on education and/or didactics with an orientation to the relevant subject. Contrary to the UK for instance, there is no competition between subject tests or test developers, as each test developer is responsible for one or sometimes more tests. This also means that the development process can differ between test developers. Since there is very little information published regarding the development of these tests, it is not possible to make any generalizations regarding test development methods and processes in detail. Still, the overall model for test development seems to be fairly similar between test developers: content experts, usually former or practicing teachers, construct the items in the tests. Thereafter, a panel of internal and external content experts review the items before and after field-testing. The standard setting is a particularly important step in the development phase, since the requirements for the test grades are determined individually for each test (more about the standard setting process below) in the different subjects, and the cut scores are determined before the tests are administered. This may seem like an unorthodox procedure, but there are reasons for this model.

The standard setting process

Due to the limited information on the test development processes among the various test developers, the following discussion will be mainly based on the tests in mathematics and science, developed at Umeå University, Sweden, a test development process currently led by one of the authors of this chapter, Anna Lind Pantzare. The development process of these tests

follows the recommendations in the Standards (AERA *et al*., 2014). There are quite elaborate blueprints defining the amount and type of items needed to measure the defined goals and knowledge requirements. The overarching ambition is, naturally, that each new test form should be parallel to the test forms previously administered. In this process a combination of item classifications, results from field trials and analyses of student work are used as indicators of parallelism. The idea is that the test development procedure will result in test forms that have similar cut scores, so that the standard setting procedure should result in only minor adjustments, if any, in relation to the intended cut scores. Ever since the implementation of the criterion-referenced grading system and the introduction of the national tests, the cut scores have been established via standard setting before test administration. The main argument given for this model is that it will prevent teachers from interpreting the test scores in a relative manner – that is, to grade on the curve.

It is well known that standard setting procedures must be implemented in a sound way to yield valid cut scores. Generally, the procedure follows these steps: selection of a representative panel; the choice of a suitable method; preparation of performance level descriptions; training of the participants to use the selected method; collection of the first round of ratings; discussion of the ratings and providing panellists with supplementary information (e.g. empirical item data); collection of one, possibly two, round(s) of reviewed ratings; evaluation of the standard setting process; and documentation of the process (Cizek and Bunch, 2007; Hambleton and Pitoniak, 2006). Moreover, this part in the development of the national tests is relying on collaborative work with experienced teachers and content experts. Normally 10–15 panellists are used in each group. The panellists are supposed to contribute with different characteristics – that is, age, gender, school size representation, experience of working with different kind of students, teaching experience and geographical differences.

The Swedish implementation of standard setting procedures follows the approach recommended in the literature except for one alteration: it does not include a separate step for the determination of performance-level descriptors. This is because the syllabus defines the knowledge requirements for each subject, and these are used as the performance-level descriptors. Since teachers regularly work with these knowledge requirements when they teach, assess and grade their students, they are supposed to be well acquainted with them. The teachers' grading experiences allow them to identify the group of borderline examinees at each grade level, which is essential in the standard setting procedure. Therefore, it has been seen as

logical to use teachers as panellists in the standard setting panels. Research has shown that, at least for the mathematics tests, since there are rather small variations in the distributions between panellists, the final decision making is rather simple (Lind Pantzare, 2017).

When it comes to method, several different standard setting methods are used, since there are many different national tests containing different kinds of items, from reading comprehension assessed by multiple choice items to mathematics items demanding a complete solution, as well as oral tasks and essays in Swedish and English.

For parts of tests containing many items, the standard setting is often done with the Angoff (1971) method, which is one of the most commonly used test-centred methods from an international perspective. While the original method was designed for dichotomously scored items, Hambleton and Plake (1995) extended the method to also include polytomously scored items. This modified Angoff method is used in establishing the cut scores for several parts of the Swedish national tests since it (1) has the capability to handle both dichotomously and polytomously scored items and (2) offers the possibility to establish cut scores before test administration. The Angoff method is one of the few methods that have both of these attributes. For the parts in the Swedish and English tests where the students are to produce longer written essays, the most common standard setting methods are variations of the Bookmark method.

The standard setting meetings follow a strict agenda: before or at the beginning of the meeting, all of the panellists receive a copy of the test form and the mark scheme. The panellists are instructed to thoroughly work through the material. When the panellists in each panel have gathered, they start the meeting by discussing the test form, as well as demands for the mark scheme in relation to the knowledge requirements. Next, the chair introduces the method that is to be used. Thereafter, a first round of individual item estimations is carried out. These estimations serve as a basis for discussions regarding the interpretations of the knowledge requirements. There is a special focus on items with large variation in the estimated item difficulties. After this discussion, a second and final round of estimations for the different grades is collected. It is very uncommon that items are deleted at this stage, and no items have been deleted after the administration of the test.

The estimates from the standard setting sessions are handled in different ways. For some of the tests the individual estimates are discussed among the panellists and, after also taking field test data into consideration, they reach consensus about the final cut scores. For some of the tests there

are two separate standard setting groups estimating the item difficulty. The final cut scores are decided within the group of test developers.

Discussion

It can probably be concluded from this chapter that the Swedish criterion-referenced education and assessment system has both strengths and weaknesses. The term 'standards-based (accountability)' has been used to describe the Swedish model (Eklöf *et al.*, 2009), since the education and assessment system is expected to evolve around the National Curriculum, and in this way strives toward giving all students an equal education, where they are assessed and graded in a valid and reliable way. However, considering the problems with maintaining standards between schools and over time, this label may be questionable since there are many aspects of non-standardization. The expected advantage of a criterion-referenced approach is that the outcome should be able to be used for giving feedback to students and parents concerning performance in relation to standardized objectives, and for monitoring educational progress in general. Both of these approaches have proved somewhat problematic, providing information with validity problems. From the perspective of educational feedback, the differences in teachers' interpretations of performance descriptors and their assessment methods may have fewer consequences for students than for teachers and schools. Students may be assessed in a strict or more lenient way, which may seem unfair and, in extreme cases, may also affect motivation and future study choices etc., but feedback is still possible – in relation to the specific interpretation of the criteria. However, there are other aspects that are problematic on a higher and on an aggregate level. School performances will be difficult to evaluate correctly since a school with more generous grading is easily mistaken for providing better education than a school that is more restrictive. The reasons for variations in grading can be many. It can be caused by teachers' different interpretations of criteria and guidelines. However, besides the difficulty of making valid and reliable assessment based on performance descriptors, it has been shown that teachers often are pressured by different stakeholders (students, parents, school leaders) for lenient grading, and particularly so among schools that are facing competition (Lärarnas Riksförbund, 2011; Wikström, 2005). This is of course serious in a system with school competition and where schools and teachers are being held accountable for their performances, while the consequences for the students are more serious when their grades are used in the selection to higher education.

The national tests have important tasks to fulfil, especially when it comes to giving teachers information on their students' performances, but perhaps especially to provide information on what type of knowledge and skills are required for the different grade levels, and thus hindering grade inflation and other unwanted variations in grading. Research has shown that the presence of national tests seems to serve this purpose, at least to some degree (Wikström, 2005). Furthermore, the national tests were initially not intended to be used as high stakes instruments, and are not designed as such. Neither were they designed for making comparisons over time, which makes the balance with test interpretation delicate: although they are the only instruments available for the purposes that are attached, they are not to be seen as examination tests, and not for strict comparisons. There have been discussions regarding whether the national tests can be adjusted to better work for school evaluations, also over time, or if other tests should be developed for this purpose. Currently a new and rather detailed framework has been proposed for how the national tests are to be developed, interpreted and used to increase their reliability and validity, and to ensure that correct interpretations are made, following a proposition by the Swedish government. To what degree test purposes and test designs are to be changed for the future is yet to be seen.

References

AERA (American Educational Research Association), APA (American Psychological Association) and NCME (National Council on Measurement in Education) (2014) *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Andersson, H. (1991) *De relativa betygens uppgång och fall*. Pedagogiska institutionen, Umeå University. Online. https://goo.gl/MM8zzi (accessed 19 June 2018).

Angoff, W.H. (1971) 'Scales, norms and equivalent scores'. In Thorndike, R.L. (ed.) *Educational Measurement*. 2nd ed. Washington, DC: American Council of Education, 508–600.

Blanchenay, P., Burns, T. and Köster, F. (2014) *Shifting Responsibilities – 20 Years of Education Devolution in Sweden: A governing complex education systems case study*. OECD Education Working Papers 104. Paris: OECD Publishing. Online. http://dx.doi.org/10.1787/5jz2jg1rqrd7-en (accessed 19 June 2018).

Cizek, G.J. and Bunch, M.B. (2007) *Standard Setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.

Erickson, G. (2017) 'Experiences with standards and criteria in Sweden'. In Blömeke, S. and Gustafsson, J.-E. (eds) *Standard Setting in Education*. Cham: Springer, 123–42.

Eklöf, H., Andersson, E. and Wikström, C. (2009) 'The concept of accountability in education: Does the Swedish school system apply?'. *CADMO*, 17 (2), 55–66.

Gustafsson, J-E. and Erickson, G. (2013) 'To trust or not to trust? Teacher marking versus external marking of national tests'. *Educational Assessment, Evaluation and Accountability*, 25 (1), 69–87.

Hambleton, R.K. and Pitoniak, M. (2006) 'Setting performance standards'. In Brennan, R.L. (ed.) *Educational Measurement*. 4th ed. Westport, CT: Praeger, 433–70.

Hambleton, R.K. and Plake, B.S. (1995) 'Using an extended Angoff procedure to set standards on complex performance assessments'. *Applied Measurement in Education*, 8 (1), 41–55.

Klapp Lekholm, A. (2008) 'Grades and grade assignment: Effects of student and school characteristics'. Doctoral thesis, Gothenburg University. Online. http://hdl.handle.net/2077/18673 (accessed 19 June 2018).

Lind Pantzare, A. (2017) 'Validating standard setting: Comparing judgmental and statistical linking'. In Blömeke, S. and Gustafsson, J.-E. (eds) *Standard Setting in Education*. Cham: Springer, 143–60.

Lundahl, L. (2002) 'From centralisation to decentralisation: Governance of education in Sweden'. *European Educational Research Journal*, 1 (4), 625–36.

Lärarnas Riksförbund (2011) *Betygsättning under påverkan*. Stockholm: Lärarnas Riksförbund. Online. https://goo.gl/5q62zr (accessed 19 June 2018).

NAE (National Agency for Education) (2014) *Sambedömning i skolan – exempel och forskning*. Stockholm: Fritzes Förlag.

NAE (National Agency for Education) (2016) *Utvärdering av den nya betygsskalan samt kunskapskravens utformning*. Stockholm: Skolverket. Online. https://goo.gl/9DE9kp (accessed 19 June 2018).

OECD (Organisation for Economic Co-operation and Development) (2015) *Improving Schools in Sweden: An OECD perspective*. Online. www.oecd.org/edu/school/Improving-Schools-in-Sweden.pdf (accessed 19 June 2018).

Parding, K. (2011) 'Forskning om den svenska friskolereformens effekter – en litteraturöversikt'. *Didaktisk Tidskrift*, 20 (4), 231–48.

SFS (2010) *800: Skollag* [School law]. Stockholm: Utbildningsdepartementet. Online. https://goo.gl/q8TWmB (accessed 19 June 2018).

Skolinspektionen (2011) *Lika eller olika? Omrättning av nationella prov i grundskolan och gymnasieskolan*. Stockholm: Skolinspektionen.

Swedish Standards Institute (2016) Online. http://www.sis.se/tema/iforvaltning/Guide/ (accessed 27 July 2018).

Tholin, J. (2006) 'Att kunna klara sig i ökänd natur: En studie av betyg och betygskriterier – historiska betingelser och implementering av ett nytt system'. Doctoral thesis, Borås University. Online. https://goo.gl/iytZaB (accessed 19 June 2018).

Vallberg Roth, A.-C., Gunnemyr, P., Londos, M. and Lundahl, B. (2016) *Lärares förtrogenhet med betygssättning*. Malmö: Malmö University. Online. https://goo.gl/dbxqU1 (accessed 19 June 2018).

Wedman, I. (1983) *Den eviga betygsfrågan: Historiskt och aktuellt om betygsättningen i skolan*. Report 48. Stockholm: Skolöverstyrelsen.

Wedman, I. (2000) *Behörighet, rekrytering och urval: Om övergången från gymnasieskola till högskola*. Report 6 AR. Stockholm: Högskoleverket.

Wedman, J. (2017) 'Theory and validity evidence for a large-scale test for selection to higher education'. Doctoral thesis, Umeå University. Online. https://goo.gl/mJeChf (accessed 19 June 2018).

Wikström, C. (2005) 'Criterion-referenced measurement for educational evaluation and selection'. Doctoral thesis, Umeå University. Online. https://goo.gl/me842x (accessed 19 June 2018).

Standardization and variability

Gudrun Erickson

Wikström's and Lind Pantzare's chapter deals with a number of critical issues in the Swedish educational system. The text is broad but manages to focus on essential aspects and implications, in particular regarding national tests and standard setting. In the following, some additional comments will be given on these two phenomena.

As shown in the chapter, there is a long tradition of national tests in Sweden, and a strong trust in teachers' responsibility for rating and grading, the latter, however, questioned and partly challenged in recent years. Furthermore, the delegation of test development to university departments, an arrangement in existence since the 1980s, is well established and largely appreciated by different stakeholders. However, external as well as internal investigations have highlighted variability at different levels as a distinct problem in a large-scale assessment system with explicit aims to strengthen individual fairness and overall equity.

Several aspects of the system demonstrate clear differences in interpretations, processes, products, outcomes and use, as shown, for example, in reports from the university groups engaged in test development. One example concerns the assignment, and how the wording in the subject syllabi should be interpreted and operationalized in tasks and tests, and to what extent standardized procedures and empirical evidence, alongside assumed impact and exemplarity, should be taken into account. Is it, for example, at all possible to assess students' ability of reasoning and analysis using closed test formats? To what extent and with what quality demands should performance-based tasks be used, given the fact that several studies have revealed alarmingly low inter-rater consistency.

The test development process is another source of variability. There is a fair amount of consensus regarding the value of collaboration between different stakeholders and the necessity of piloting materials. However, the selection, as well as the number, of test-takers in piloting and pre-testing varies considerably, as does the use of anchor items for test equating purposes. In addition, there are considerable differences in analytical methodology, in particular regarding the perceived value and use of quantitatively oriented procedures. The number of items and tasks

also vary from one test to another. Furthermore, there is evident variability regarding standard setting; mostly, basic Angoff-related procedures are used, but differences are large when it comes to the type of evidence used to arrive at the final recommendation for cut scores and benchmarks.

Finally, variability is evident also in the overall stability of different subject tests over time and in teachers' use of aggregated national test grades when awarding final grades. As a result, several measures have been taken to strengthen the system, for example a decision by the government about an extensive inquiry regarding the future of the national tests, and a proposed framework for the assessment system, developed academically on commission by the National Agency for Education. The latter has recently been delivered and is currently in the process of analysis by the NAE and the different test development groups. Quite predictably, reactions have been mixed. Attempts to increase validity, reliability and stability, for example through a certain degree of standardization of processes and products, are not generally approved of. A number of important decisions remain to be taken, and to what degree positive effects will be achieved is something yet to be seen.

From cohort-referencing to criterion-referenced grades in Sweden

Jan-Eric Gustafsson

The chapter on Sweden provides a comprehensive description and discussion of essential aspects of the systems for student assessment and grading. In this commentary I have chosen to take a historical perspective, focusing on the transition from cohort-referenced to criterion-referenced grading.

One of the most important texts on assessment and grading ever published in Sweden is the commission report SOU 1942:11. This report proposed a system of grading in compulsory school designed to yield equitable and comparable teacher-assigned grades. A criterion-referenced system was considered by the commission, but it was rejected with reference to the fact that verbally formulated criteria cannot achieve the degree of precision required for purposes of grading. Instead a cohort-referenced system was proposed which was designed to give ample room for teacher assessments. The proposed system was based on research showing that teachers are highly skilled in evaluating the relative merits of their own students, but that they cannot compare the achievements of their own students with those of students in other classrooms. It was therefore suggested that so-called 'standard tests' should be developed to supply information about class-level performance. Along with an assumption about a normally distributed population, this information would be sufficient for the teacher to know approximately how many grades at different levels would be available for the class. While student performance on the standard test was to be taken into account when grading individual students, the teacher was also expected to bring in information from other sources in the assessment, such as classroom performance and results on teacher-made tests. This cohort-referenced, so-called 'relative' grading system was implemented in compulsory school in the early 1950s, and later on it was also implemented in upper-secondary school.

Within the framework of this robust system, Swedish teachers were granted wide responsibilities for grading for high stakes purposes. However, there also was criticism of the relative grading system. One complaint was that the mean level of achievement of the population was assumed to be

constant from one year to another, giving the impression that there could be no improvement of achievement. Another point of criticism was that the system encouraged competition rather than collaboration.

These, and other criticisms, are likely causes of a sudden decision to discontinue the relative grading system in conjunction with the introduction of new curricula in the 1990s, and it was replaced by a criterion-referenced 'goal- and knowledge-related' grading system. This grading system was planned to serve three functions. The first was reliable grading at the individual level, the second was to serve purposes of evaluation at intermediate levels of the school system (e.g. school, municipality) and the third was to provide information about development of levels of achievement at national level over time. The idea to use the grades for these multiple purposes was based on the assumption that the grades would provide unbiased information about the extent to which the different goals had been reached. However, this assumption proved false, and the three planned functions have not been adequately implemented.

It soon became clear that there were large differences among teachers and schools in leniency of grading and also that there was a substantial grade inflation. National tests were expected to support equitable grading. However, it was never made clear in what way or to what extent the national test results should influence the grades of individual students. Furthermore, in most subjects the tests were designed as performance tests, requiring students to produce large amounts of text or other output, which had to be interpreted and assessed by the teachers. However, the Swedish Schools Inspectorate concluded that the teacher ratings were unreliable and biased in favour of the teachers' own students, leading to blaming and shaming of the Swedish teachers in the media (Gustafsson and Erickson, 2013).

The national tests also suffer from the problem that the national averages in most cases vary substantially from one year to another, while the grades tend to increase over time. In contrast, in the international comparative assessments, the results for Sweden have been declining since the mid-1990s. The criterion-referenced grades thus cannot be used for purposes of national assessment of achievement trends. The poor measurement characteristics of the criterion-referenced grades also cause relations with other variables, such as family background, to be underestimated, giving the false impression that equity of schooling outcomes has improved over time (Gustafsson and Yang Hansen, 2017).

Thus, there have been a large number of negative consequences of the introduction of the criterion-referenced grading system. One source of these negative consequences is the inherent impossibility to formulate goals

and criteria in sufficiently precise terms to achieve comparability of grading, and another main source is the lack of a system of national tests that can provide teachers with the necessary support in their work with assessment and grading.

References

Gustafsson, J.-E. and Erickson, G. (2013) 'To trust or not to trust? Teacher marking versus external marking of national tests'. *Educational Assessment, Evaluation and Accountability*, 25 (1), 69–87.

Gustafsson, J.E. and Yang Hansen, K. (2017) 'Changes in the impact of family education on student educational achievement in Sweden 1988–2014'. *Scandinavian Journal of Educational Research*. Online. www.tandfonline.com/doi/full/10.1080/00313831.2017.1306799 (requires subscription).

SOU 1942:11. *Betänkande med utredning och förslag angående betygssättningen i folkskolan*. Stockholm: Ecklesiastikdepartementet. Online. https://goo.gl/6P5nVY (accessed 19 June 2018).

Chapter 13

Setting Standards in the United States: The Advanced Placement programme

Deanna L. Morgan

Introduction

The College Board is a mission-driven, not-for-profit organization that connects students to college success and opportunity. Founded in 1900, the College Board was created to expand access to higher education. Today, the membership organization is made up of over 6,000 of the world's leading educational institutions and is dedicated to promoting excellence and equity in education. Each year, the College Board helps more than seven million students prepare for a successful transition to college through programmes and services in college readiness and college success – including the SAT and the Advanced Placement programme (AP). The organization also serves the education community through research and advocacy on behalf of students, educators and schools. The College Board is headquartered in New York but has regional offices across the United States and in Puerto Rico.

In the United States the education field has consistently moved towards standards-based testing. As a result, the need to quantify when a student has shown sufficient knowledge and skills on a set of content standards has evolved. While this movement has been in place for a number of years, the 2001 Elementary and Secondary Education Act (ESEA) legislation, also known as the No Child Left Behind Act (NCLB), has played an important role in accelerating the movement and focusing attention on standards-based assessment. NCLB legislation required that all states have a standards-based test in grades 3 through 8 in reading and mathematics and that all students are tested. The standards-based test must, at the least, identify students as basic, proficient or advanced according to the individual state's content standards. Additionally, states must assess all students, including those with significant cognitive disabilities.

The NCLB Act redefined the role of the US federal government in primary and secondary education. Along with mandating annual student testing in Grades 3–8, it stipulated that assessments provide adaptations and accommodations for students with disabilities as defined in the Individuals with Disabilities Acts of 1991 and 1997. It also mandated the reporting of assessment results and state progress by student groups based on socio-economic status, race and ethnicity, disability status and limited English proficiency. However, it is important to note that the mandate did not include a common curriculum or assessment, making it very difficult to compare performance across states. The ultimate goal was for all students to reach proficiency by the year 2014. As the United States approached this deadline, it became clear that this ultimate goal would not be met, and states joined to form consortiums to develop and administer common assessments within each consortium measuring the Common Core State Standards (Common Core). The Common Core places an emphasis on defining content and performance standards indicative of college and career readiness (Morgan and Perie, 2013).

Currently, however, a large number of states have withdrawn from the consortiums and the Common Core curriculum. As such, states and in some cases districts or even individual schools have the ability to define the curriculum that will be taught and to choose the assessments that will be offered. This great diversity of options makes it increasingly difficult to measure comparable student performance. Efforts have been made to use performance on the National Assessment of Educational Progress (NAEP) as a baseline and compare state performance relative to their NAEP performance. This is problematic, however, due to small sample sizes, matrix sampling of content offered on the exams, and low student motivation since no consequences for the student are tied to the test result. In 2015, the Every Student Succeeds Act (ESSA) was signed reauthorizing the ESEA and revising many parts of the law under NCLB but still failing to mandate a common curriculum or common assessment. The College Board holds a unique position in that the same curriculum and assessments are used across the United States and in other countries around the world with no government oversight or accountability. Accountability of the College Board and the AP programme is to the members of the organization, users of the product (both schools and students), and higher education institutions where decisions may be made to accept or not accept AP scores.

The Advanced Placement programme

The College Board's AP programme provides an avenue for high school students to pursue college-level content with the potential of earning college credit, placement into a college course beyond the introductory course (advanced placement), or both at an institution of higher learning. The programme has 37 college-level courses that culminate in either a rigorous exam or a final product(s) that will be evaluated and scored. AP courses last for one school year or the equivalent: the course may last for only half of the school year if the classes are extended length such as found in some block scheduling arrangements. AP students receive a categorical score of 1 to 5 that is based on their exam performance or final product(s), through-course assessment components which are completed during the course rather than at the end, or both. Generically, the programme describes the categorical scores, or AP grades, as:

5 = extremely well qualified

4 = well qualified

3 = qualified

2 = possibly qualified

1 = no recommendation

While the AP programme recommends that students be considered for college credit or advanced placement with a score of at least 3, it is at the discretion of the individual institution whether they will accept an AP score, if the student will receive credit and/or advanced placement, and what score will be required for the credit and/or advanced placement.

The programme began in the early 1950s and grew out of five pilot projects initiated at that time by the Ford Foundation and the Carnegie Corporation in response to concerns about the need to have a rigorous education and avoid mediocrity for gifted and talented students in the post-World War II and early Cold War era. The College Board took over what remained of these efforts with continued early funding from the Ford Foundation in 1954 (Lacy, 2010). Between 2002 and the present time, AP focused heavily on reviewing and redesigning the courses and exams with the goal that courses would focus on key knowledge, skills and abilities that students should know and be able to do with an eye toward deeper understanding of fewer concepts rather than a shallower coverage of a broader array of content. A key piece of this was the implementation

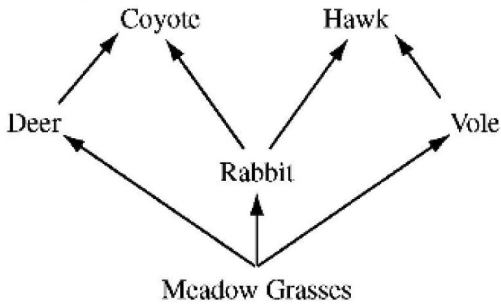
of evidence-centred design (ECD). Evidence-centred design (ECD) is an orientation towards assessment development. It differs from conventional practice in several ways: (a) the amount of work required up front in the design phase (i.e. before items are written); (b) the prioritized role of observable evidence in design and development; and (c) the documentation and use of *claims*, *evidence* and *task models* (Huff *et al*., 2010; Mislevy *et al*., 2003; Hendrickson *et al*., 2013). The review also included current psychometric practices such as how cut scores are established to set the standard for each AP grade, which equating practices are used to maintain the standards from one administration to the next and across forms, how item quality is evaluated and what range of values is acceptable, and how reader reliability is monitored for the scoring of constructed response exams. All aspects of the exam process were evaluated against the Standards for Educational and Psychological Testing (AERA, APA and NCME, 2014). A strong recommendation was made by a panel of experts to transition to panel-based standard setting for setting the AP cut scores used to assign students to the AP grades 1–5. Historically, the programme used college comparability studies to set the standards in an effort to maintain the congruence between college-level and AP assessment standards. However, problems with obtaining a representative sample of college level examinees, motivation, grade inflation, match to curriculum and timing during the year and other factors made the results at times questionable or unusable. Beginning with the AP Environmental Science Course in 2011, all AP standards are established through a panel-based standard setting.

The assessment process

Nature of assessments

Advanced Placement currently offers 37 different course and exam programmes. Exams are administered in the first two weeks of May. The majority of the exams begin with a section of multiple choice (MC) items that each have four response options (A) to (D) (see Figure 13.1), and then have multiple-select multiple-choice questions (see Figure 13.2) and short answer free-response (FR) questions (see Figure 13.3). The length of the exam sections varies as needed to cover the content and skills specific to each content area.

The following is a food web for a meadow habitat that occupies 25.6 km^2. The primary producers' biomass is uniformly distributed throughout the habitat and totals 1,500 kg/km^2.



Developers have approved a project that will permanently reduce the primary producers' biomass by 50 percent and remove all rabbits and deer.

Which of the following is the most likely result at the completion of the project?

(A) The biomass of coyotes will be 6 kg, and the biomass of hawks will be 0.5 kg.

(B) The biomass of coyotes will be dramatically reduced.

(C) The coyotes will switch prey preferences and outcompete the hawks.

(D) There will be 50 percent fewer voles and 90 percent fewer hawks.

Figure 13.1: Multiple choice item

On a day that is warm and sunny, a car is parked in a location where there is no shade. The car's windows are closed. The air inside the car becomes noticeably warmer than the air outside. Which of the following factors contribute to the higher temperature? Select two answers.

(A) Hotter air rises to the roof of the car and cooler air falls to the floor.

(B) The body of the car insulates the air inside the car.

(C) Electromagnetic radiation from the Sun enters the car and is absorbed by the materials inside.

(D) The body of the car reflects electromagnetic radiation.

Figure 13.2: Multiple select multiple choice item

Use the passage below and your knowledge of European history to answer all parts of the question that follows.

"One of the greatest afflictions of a king is when his people are torn apart, as when in one house the children against the wish of their father are banded together one against the other. . . . So the war is entirely contrary to the establishment of proper order and the increase of your grandeur. . . . Your Majesty will be aware that we by no means approve of the so-called reformed religion, but . . . the cinders of the fire of this so overwhelmed kingdom are still so hot that it is impossible to hold them in your hand without burning your fingers. . . . We beseech you, Sire, very humbly to believe that whoever desires this civil war is ungodly, and to take notice of two maxims: the first, that the peace of your subjects lies in the union of your princes; and the other, that violence eventually leads only to self-destruction."

Petition of nobles to the king of France, 1577

a) Briefly identify and describe ONE cause of the conflict discussed in the petition.

b) Briefly identify and describe ONE result of the conflict discussed in the petition.

c) Briefly identify and describe how one country in early modern Europe other than France dealt with the type of conflict discussed in the petition.

Figure 13.3: Short answer free response question

Most exams take approximately three hours with timing applied as appropriate to the section or item type. The Language and Culture exams feature both listening and speaking sections in addition to the MC and FR items. A few exams include either short answer questions, multiple select multiple choice items, or grid-in items that require the student to code a numeric response on the answer sheet. History exams offer students two long essay prompts, from which the student chooses one to respond to. Additionally, three exams have been launched that include multiple performance tasks that students prepare for during the year and submit for scoring in lieu of or in addition to sitting an exam (see Figures 13.4 to 13.7). Students in one of the three Studio Art programmes are required to submit a portfolio of work with accompanying documentation or responses regarding the work and its motivation. The curriculum and exam redesign efforts have focused more solidly on skills and greater standardization of exams and achievement level descriptors (ALDs) within subjects: for example, the three history exams have the same exam structure and the same ALDs with differing supporting examples as appropriate for the specific area of history.

AP Seminar Performance Task 1: Team Project and Presentation

Student Version

Weight: 20% of the AP Seminar score

Task Overview

You will work in teams of three to five to identify, investigate, and analyze an academic or real-world problem or issue; consider options and alternatives; and present and defend your proposed solution(s) or resolution(s). The components that comprise this task are the Individual Research Report and the Team Presentation and Defense. These components are made up of the following elements, each of which you will need to complete in order to fulfill the task requirements:

Task Elements	Length	Date Due (fill in)
Individual Research Report	1200 words	
Team Presentation	8–10 minutes	
Oral Defense (part of Team Presentation)	Each student responds to 1 question	

In all written work, you must:

- Acknowledge, attribute, and/or cite sources using in-text citations, endnotes, or footnotes, and/or through bibliographic entry. You must avoid plagiarizing (see the attached AP Capstone Policy on Plagiarism).
- Adhere to established conventions of grammar, usage, style, and mechanics.

Task Directions

1. **Team Coordination**

- **As a team,** collaborate to identify an academic or real-world problem or issue (e.g., local, national, global, academic/theoretical/philosophical).
- Develop a team research question that can be viewed from multiple perspectives.
- Conduct preliminary research to identify possible approaches, perspectives, or lenses.
- Divide responsibilities among group members for individual research that will address the team's research question.

2. **Individual Research Report** (1200 words)

- Work with your team to decide and clarify your individual approach to the team question.
- Throughout your research and as a team, continually revisit and refine your original team research question to ensure that the evidence you gather addresses your collective purpose and focus.

Figure 13.4: Through-course performance task

- **On your own,** investigate your assigned approach, range of perspectives or lens on the problem or issue of your team research question.
- Identify a variety of sources that relate to your particular approach to the team research question.
- Analyze and evaluate the relevance and credibility of sources and evidence.
- Synthesize the perspectives you have gathered and chose which ones would be most valuable to share with your team in your individual report.
- Consult with your peers to get feedback and refine your approach throughout.
- Ensure that the report that you submit is entirely your own work.
- Present your findings and analysis to your group in a well-researched and well-written report in which you:
  - Identify an area of investigation and explain its relationship to the overall problem or issue.
  - Summarize, explain, analyze and evaluate the main ideas and reasoning in the chosen sources.
  - Evaluate the credibility of chosen sources and relevance of evidence to the inquiry.
  - Identify, compare and interpret a range of perspectives about the problem or issue.
  - Cite all sources that you have used, and include a list of works cited or a bibliography.
  - Use correct grammar and style.
- Do a word count and keep under the 1200-word limit (excluding footnotes, bibliography, and text in figures or tables).
- Remove any references to your name, school, or teacher.
- Upload your document to the AP Digital Portfolio.

3. **Team Collaboration and Argument Construction**

- Read all team members' reports.
- Teach other team members what you learned so that all team members understand all perspectives presented in the reports (in the Oral Defense, you may be asked about any team member's work)
- Collaboratively synthesize and evaluate individual findings and perspectives to create a collective understanding of different approaches to the problem or issue.
- Consider potential solutions or resolutions to your team's problem or issue.
- Conduct additional research on solutions or resolutions.
- Evaluate different solutions in relation to context and complexity of the problem.
- Propose a solution or resolution to your problem or issue.
- Develop an argument to support your proposed solution.

Figure 13.5: Through-course performance task

4. **Team Multimedia Presentation and Defense** (8–10 minutes)

   Together with your team, develop a presentation that presents a convincing argument for your proposed solution or resolution. Your claims should be supported by evidence and you should show you have considered different perspectives and the limitations and implications of your proposed solution or resolution.

   When preparing your presentation:

- Develop and prepare a multimedia presentation that will present your argument for your proposed solution or resolution.
- Plan each team member's role in the presentation design and delivery.
- Design your oral presentation with supporting visual media, and consider audience, context, and purpose.
- Prepare to engage your audience using appropriate strategies (e.g., eye contact, vocal variety, expressive gestures, movement).
- Prepare notecards or an outline that you can quickly reference as you are speaking so that you can interact with supporting visuals and the audience.
- Rehearse your presentation in order to refine your design and practice your delivery.
- Check that you can do the presentation within the 8- to 10-minute time limit.
- Practice asking each other questions about the process and product of this project to prepare for your oral defense.
- Deliver an 8–10 minute multimedia presentation in which you:
  - Evaluate potential resolutions, conclusions, or solutions to problems or issues raised by different perspectives considered by your team by considering their implications and consequences.
  - Present a well-reasoned argument that links claims and evidence about why you chose your proposed solution or resolution.
  - Identify and explain objections, implications, and limitations of competing perspectives.
  - Engage the audience with an effective and clearly organized presentation design.
  - Engage the audience with effective techniques of delivery and performance.
  - Demonstrate equitable participation and engagement of all team members.
- Following the presentation, your team will defend its argument. Your teacher will ask each individual team member a question in which you will:
  - Reflect on experiences of collaborative effort and defend your team's work. Each team member should be prepared to answer questions about any part of the presentation or research process (including information that others in your team have researched and/or presented).

Figure 13.6: Through-course performance task

Sample Oral Defense Questions

Here are some examples of the types of questions your teacher might ask you during your oral defense. These are *examples only*; your teacher may ask you different questions.

1. Describe how the content of the team presentation was changed as a result of group discussion.
2. Student A, how did the group decide to include Student B's perspective/lens/ conclusions into the overall presentation?
3. Student A, give one specific way that your thinking changed as a result of learning about Student B's findings.
4. In the future, what change would you make to your group norms, and how would you expect that to improve the team presentation.
5. Reflecting on your colleagues' work, which one had the greatest impact on your overall understanding of the problem your group identified?
6. In what way did you improve your ability to work with a group as a result of this project?
7. What is an example of a compelling argument from one of your peer's individual reports that you decided to exclude from your team presentation and why?
8. What is a way in which your team's resolution makes you think differently about your own individual research?
9. What was the strongest counter argument to the solution or conclusion your team identified and why?
10. Describe an argument from one of your peer's individual reports that made you think differently about your team's solution or conclusion?
11. Having finished your project, what if anything do you consider to be a gap in your team's research that, if addressed, would make you feel more confident about your conclusion?

Figure 13.7: Through-course performance task

Examinations

The College Board owns the exams and works with curriculum experts both on staff and on committees developed to represent higher education and secondary education in that content area. Committee members and content experts from both the College Board and the Educational Testing Service (ETS) work collaboratively to define content and skills, write items, review item performance data, create scoring rubrics and participate along with many other content experts in the annual reading in June to read and score student responses to the free response questions and performance tasks. Although the College Board is the final decision maker about exam content and design, external expertise from stakeholders is an integral part in the process. Routinely, external content experts are invited to provide feedback through surveys, participation on committees and a variety of other routes. All work is reviewed by College Board and ETS for accuracy and bias with

many levels of security safeguards in place to protect confidential items and materials including student confidentiality in the handling of personally identifying information. Embedded pre-testing is implemented in some AP exams to increase item quality by examining item performance prior to being used to contribute to a student's score. However, not all AP exams use embedded pre-testing due to concerns about increasing test length and security.

School-based assessment (coursework)

For the majority of the courses and exams, the only prescribed activity is the exam in May. However, AP teachers may offer a variety of learning opportunities and activities at their discretion to assist student learning and result in a reportable classroom grade for the purpose of the high school experience and student record. The exceptions are the courses that feature performance tasks or portfolios that are submitted and graded in addition to or in lieu of an end of course exam. These courses may require an in-class presentation or oral defence that is graded by the teacher according to the scoring rubrics, on which the teacher must be trained. Some written products may also be scored by the teacher but are then also scored at the official reading in June by an independent rater. Presentations are not scored at the reading and receive the score provided by the teacher. Currently, no moderation is done for scores that are only assigned by the teacher.

Marking completed examination papers

With the exception of the course with a performance task component mentioned above that is scored by the teacher, all MC questions and questions that are able to be gridded on an answer sheet are scanned and scored electronically, generally within two weeks or less from the exam date as materials are returned to ETS by the schools where the exams were administered. When materials are returned, the answer sheets are separated from the students' response booklet for the free response questions. The free response question booklets are then sorted and bundled for processing in preparation for the annual reading, which typically occurs in the first two weeks of June. Each AP subject is assigned to one of multiple reading sites around the United States. Secondary teachers and higher education faculty are recruited to participate in the reading of responses for a specific subject area and spend approximately seven days in a large convention centre scoring the written or digital responses that have been received. Each reading begins with training on pre-selected samples and a thorough review of the rubric for the specific question each reader has been assigned

to score. A reader may score more than one question during the week but focuses on only one question at a time and will be retrained on the new question before starting to score again. AP-constructed response questions are scored by a single reader with periodic back reads by table readers to ensure quality in addition to the presence of calibration papers, which are circulated throughout the process to verify that a reader is aligned with the rubric. Calibration papers have known scores and readers must score the papers correctly to continue reading, or they receive additional training on the rubrics before being allowed to resume scoring. Reader reliability studies are conducted frequently at the AP readings to obtain double score data on student responses for a sample of the population testing and allow for reader agreement rates and generalizability results to be produced as a measure of reader quality and consistency. Along with accomplishing the work of scoring the responses, the AP Program has found that the reading is an excellent professional development opportunity for educators, and many of the subject areas have developed unique cultures that continue from year to year as readers return and new readers are added. Though many efforts are underway to lower costs and be more efficient through the introduction of online distributed scoring, it is unlikely that the readings will ever be completely replaced due to the other benefits and goodwill derived by the gathering of so many educators in one place for a common purpose.

Standard setting process

Determining grades

Since 2011 the AP programme has used panel-based standard setting to make recommendations about cut-score locations. Fifteen subject matter experts (SMEs) in the subject of the exam are convened. Seven are teachers of the AP course and the remaining eight are higher education teachers of a comparable college level course for which students earning an acceptable score on the exam may receive credit. In addition to expertise and experience, SMEs are also selected to represent a diverse group of gender, race/ethnicity, geographical location and years of experience teaching, with additional considerations as needed depending on the specific needs of the subject, for example both heritage and non-heritage speakers for the Language and Culture exams. At the beginning of the study, the SMEs are asked to complete a biographical data form for use in summarizing panellist characteristics, and evaluation-form data is collected throughout the standard setting meeting as evidence of procedural validity (Kane, 2001; Hambleton *et al*., 2012; Pitoniak and Morgan, 2012, 2017). Additionally, panellists are required to sign a confidentiality form since they are

working with sensitive and confidential test materials during the standard setting process.

Initially, SMEs receive an overview of the course and exam, the AP grades (1–5), and a brief introduction to the purpose of the meeting. This is followed by the opportunity to experience the exam that will provide the SMEs with a frame of reference for considering student performance in the context of the entire exam and under conditions close to those encountered operationally. SMEs do not have access to answer keys during the exam experience. This activity familiarizes SMEs with the exam questions, as well as with the rigour and time constraints experienced by students who take the exam. Following completion of the exam, an answer key and analytic rubric are provided to SMEs so they can score their own performance.

SMEs then have an opportunity to review and discuss items they found especially difficult or confusing with an emphasis on characteristics of specific items, what is being measured and what factors make the item easy, difficult, or perhaps confusing. Any comments of a critical nature or editorial type beyond the scope of the standard setting task are collected and shared with the appropriate person(s) in the AP programme for later follow up.

Historically, prior to 2011, standards were set on the AP exams through a college comparability study that included giving the exam to college-level students enrolled in the relevant course for the exam and mapping their exam performance to their expected grade in the course, resulting in the following relationship:

AP Grade 5 = A or A+ in the corresponding college course

AP Grade 4 = B, B+, or A- in the corresponding college course

AP Grade 3 = C, C+, or B- in the corresponding college course

AP Grade 2 = D, D+, or C- in the corresponding college course

AP Grade 1 = Grade below D in the corresponding college course

While the AP programme has moved to panel-based standard setting, a desire remained to have some grounding in college student performance considered by the SMEs. As a result, the higher education SMEs are asked to administer a comparable but shortened version of the exam to their students, to provide the results (along with expected course grades) to the College Board in advance of the meeting, so analyses can be completed to share at the standard-setting meeting for comparison with the performance

of the AP student population taking the exam. These results are shared with the SMEs following the exam experience and then revisited during the meeting as other results become available. However, it should be noted that problems which plagued the college comparability studies used prior to 2011, and led to the recommendation to move to panel-based standard setting, also impact the results of this mini-comparability (mini-comp) study for the standard-setting meeting, limiting its value. College comparability study results were often found to be unusable due to many limitations. One of the biggest issues facing college comparability studies was the recruitment of large samples of students and institutions to participate in the study. Recruitment, always difficult, often suffered from a large rate of attrition. Additionally, the students may not be sufficiently motivated to try their best on the exam, and results were often accompanied by notes from the professors that attested to that. The exams administered in the studies were not always fully representative of actual AP exams due to time limitations to fit within the class time of a college class. Because exams were administered at multiple sites by professors and then scored by those professors, a lack of standardization across administrations and scoring practices further affected the data used in the studies. Also, participating schools may not be representative of the colleges that accept AP scores. The resulting data, which may be limited by one or more of these issues or others not included, did not provide a high degree of confidence in the fidelity of the results. The final piece of the comparability study involved mapping college student performance on the exam to the expected class grade provided by the professor. Class grades are notorious for inflation due to characteristics other than knowledge and ability in the subject area, such as participation, attendance, politeness and a variety of other personal characteristics that may cause the professor assigning grades to be more generous or offer the benefit of doubt to a student and raise their grade even when academic performance may not warrant it. The additional complication of lack of standardization in the assignment of class grades from one professor to another when combined with the two pieces of data used in the study (study exam data and expected grades) which are also less than stellar, makes the continued use of comparability studies as the primary basis for setting performance standards untenable.

Following the presentation of the mini-comp results, the SMEs are introduced to a set of ALDs specifically designed for the course and exam on which they are working. ALDs describe the borderline knowledge, skills and abilities that are required for a student to be placed into each AP grade category, also referenced in some publications as the minimally

competent examinee. Discussion of ALDs prior to assigning standard setting ratings helps to establish a common understanding across standard-setting panellists of the meaning of the borderline of each score category in terms of what students at the borderline of each category know and are able to do. In essence, the ALDs serve as benchmarks, or anchors, during the rating task.

A variety of methods have been proposed for setting performance standards on educational assessments. Despite procedural similarity across many panel-based standard-setting techniques (Hambleton *et al*., 2012), Cizek (2012) describes at least ten separate standard-setting processes with a host of modifications that yield even more methods that can be used to collect ratings from panellists. In spite of the numerous methods, various modifications described as Angoff standard-setting procedures remain among the most widely used (Angoff, 1971; Plake and Cizek, 2012). It should be noted that the Angoff methods derive from a brief description and footnote in the second edition of *Educational Measurement* and are typically not implemented as originally described; thus most of the methods are more accurately referenced as Modified Angoff methods. The Angoff method and its variations are criterion-referenced standard-setting methods that require panellists to estimate the probability that a 'minimally acceptable person' (i.e. a borderline examinee) will answer an item correctly. These probabilities are then summed to produce recommended cut scores. A Modified Angoff standard setting method (Plake and Cizek, 2012) is used to collect SME ratings for the majority of the AP exams, which include a combination of multiple choice and free response items with a variation known as Mean Estimation used for the free response items. Recently, exams with only free response items have been launched by the AP programme, and these exams utilize a different panel-based methodology. Criterion-referenced methods that require panellists to make judgements in reference to a set criterion (the ALDs that define the knowledge, skills and abilities of the borderline examinee at each cut score) are preferred because it is critically important that students earning each of the AP grades 1–5 be able to demonstrate the knowledge, skills and abilities described. Other norm-referenced methods, such as comparability studies, which focus primarily on the percentage of test completion as it relates to the performance of a norm group on taking the test (college students taking the AP exam) and establishes cut scores numerically or based on classroom grades as opposed to in relation to knowledge of the subject matter, are likely to result in poor validity when student grades and readiness for the next course are evaluated.

SMEs receive training on the concept of the borderline examinee, using the ALDs, the Modified Angoff Method and Mean Estimation, including the process to follow to make ratings on items using both methods and a chance to practice the process with a select sample of items prior to doing any ratings that will contribute to the final recommendation. The training includes the opportunity for questions and discussion, and SMEs are asked to complete an evaluation form at the conclusion to indicate their level of understanding and readiness to proceed to the real task. The task that each SME must complete requires they use the ALDs to represent the borderline examinee in each of the AP grade categories (1–5), then provide an expected probability for correctly answering each of the multiple choice items, and estimate the mean score of 100 borderline examinees at each achievement level in the rubric for the free response items. In order to ease the cognitive demand during rating, panellists are asked to imagine a group of 100 borderline students in each AP grade category, and estimate the number who would correctly answer each multiple choice item. For the free response items, panellists are asked to estimate the average score on the rubric that those same 100 students would receive on each item. Students in borderline groups are described in terms of 'cuts' that distinguish between AP grade categories. These groups are described as follows:

- *Examinees at the 4/5 cut:* borderline examinees who receive a score of '5' on the AP exam; these students represent a minimally qualified examinee for the AP grade of '5'
- *Examinees at the 3/4 cut:* borderline examinees who receive a score of '4' on the AP exam; these students represent a minimally qualified examinee for the AP grade of '4'
- *Examinees at the 2/3 cut:* borderline examinees who receive a score of '3' on the AP exam; these students represent a minimally qualified examinee for the AP grade of '3'
- *Examinees at the 1/2 cut:* borderline examinees who receive a score of '2' on the AP exam; these students represent a minimally qualified examinee for the AP grade of '2'

SMEs are restricted to expected probability ratings between 20 and 95 in intervals of five for the multiple choice items. They are not allowed to provide ratings below 20 in order to prevent a cut score that would allow a student to receive a grade above AP 1 by guessing due to chance. Similarly, SMEs are not allowed to provide ratings greater than 95 in recognition that perfect performance is not common, nor a reasonable expectation of the borderline examinees. Additionally, this helps control for examinees

being required to earn a perfect score to be placed into the highest score category. For the free response items, SME ratings are restricted to average rubric scores between 0 and the maximum score on the rubric in intervals of 0.5. Because four cut scores are needed to assign the five AP grades, SMEs provide ratings for the four borderline groups simultaneously on each item. SMEs provide ratings by entering the appropriate value into a googledocs spreadsheet by typing the value or using a pull-down menu. Spreadsheets are constrained so that only valid values are available for use by the SME. During training, SMEs are asked to share their ratings for specific items and the rationale for how they came to that rating. A variety of ratings for an item is not uncommon, and SMEs are informed that consensus is not a goal of the meeting; it is expected that the variance among ratings will decrease as the meeting progresses, but it is not expected that everyone will be in exact agreement.

After completion of the previously mentioned evaluation form and a brief review of the forms to ensure no further training is necessary, the SMEs provide two rounds of ratings with discussion and feedback provided between rounds. Following Round 1 of ratings, SMEs complete an evaluation form that provides another opportunity for the facilitator to ensure no additional training is needed and provides further evidence for the procedural validity of the process. When all SMEs have submitted ratings, feedback is provided to each in the form of the median rating of the group on each item, which can be compared to their own rating, and the difficulty of the item based on actual student performance in the form of the percentage of examinees answering the item correctly. The percentage correct provides the SMEs with a reality check for consideration if the expected probabilities they are assigning are drastically different from how examinees actually performed. SMEs are divided into small groups of four to five people and encouraged to compare individual item ratings and discuss rationales for those ratings. After the small group discussions, the larger group reconvenes and discusses highlights from the small groups so that everyone is on the same page. It is not expected that any SME will alter their ratings as a result of the discussion, but it is common that a previously unconsidered perspective is shared during the discussion that will result in one or more changes to the Round 1 ratings. The results of the AP Teacher Survey are also shared as part of the discussion of the free response items. The Teacher Survey asks AP teachers for that subject area to provide the number of rubric points they would expect an examinee to earn on each free response item to earn an AP 3 and to earn an AP 5. Teachers completing the survey have the free response Items, Rubrics and ALDs available for use

when providing their estimates. This information is used as a reality check for the standard setting panel to compare against the number of required points based on the Round 1 ratings. Before SMEs begin their individual work rating items for Round 2, impact data is shared in the form of the expected distribution of students earning each AP grade if the Round 1 results remained intact as the recommended standards to the AP Program. The Round 1 recommendations are also applied to the data from the mini-comp study to produce a distribution of expected performance for that population as well.

During Round 2, SMEs are instructed to review each item to confirm their rating provided in Round 1 or to provide new ratings as they deem appropriate based on the information that was presented during the discussion. After all SMEs have submitted their Round 2 ratings, the updated impact data based on Round 2 ratings is shared, and the SMEs have an opportunity to discuss their viewpoints on the reasonableness of the recommended standard following Round 2 ratings. This discussion is often quite lengthy and is very informative to the AP Program staff observing the meeting. Many of the points from this discussion are recalled and considered later by the AP Program in deciding the final cut scores that will be adopted and applied operationally. The variance of the judgements, Standard Error of Judgment (SEJ), is calculated after each round of ratings, and the expectation is that this value will be relatively small and will decrease in Round 2. If SMEs seem unhappy with the results after Round 2 and/or the SEJ indicates that agreement has decreased resulting in a larger SEJ in Round 2 than in Round 1, a third round of ratings may be collected using the same method or, when appropriate, another method, such as a survey of each SME's minimum and maximum acceptable number of points on the test holistically, as a compromise method to provide another data point for decision making. Following the final round of discussion, panellists are asked to complete a final evaluation form to provide additional evidence of the procedural validity of the standard setting meeting and share any feedback they have about the process, facilities, results or any other topic desired.

After the Modified Angoff standard setting method has been used for an exam, in almost all subject areas, AP grade standards are subsequently maintained across administrations and forms through Common-Item Non-Equivalent Groups equating (Kolen and Brennan, 2004).

Variation for all free response exams

Two newly launched AP exams, AP Capstone Seminar and AP Capstone Research, are composed of only performance tasks and do not lend themselves well to the rating task described above. The process described is the same; only the rating task differs. The Performance Profile Method (Morgan, 2004) is used to make cut score recommendations on these assessments. Using real student performance data, a profile of performance by an individual student on all performance tasks is created. A set of approximately 60 of such profiles that span the range of performance from very low to very high is developed by selecting from the most frequently earned profile combinations, a representative student performance at each total score point. The profile set is randomly ordered, and the SMEs review each profile against the ALDs and make an assignment of the profile into one of the five AP performance levels. When all SMEs have made an assignment into an AP performance level for each profile in the set, the assignments are summarized to show the frequency with which each profile was assigned to each performance level. Profiles assigned to multiple performance levels are discussed by the group, and SMEs are given an opportunity to make adjustments to the assignments. After Round 2 of the assignments, impact data is shared in the form of the expected distribution of students in each AP grade if the Round 2 assignments remained in place. SMEs then have a third opportunity to make adjustments to the profile assignments, and the results from Round 3 become the recommended standard to the AP programme.

Political and public controversies and debates with the AP programme

The AP programme is highly regarded for its rigorous curriculum and examination. The number of AP exams that a student takes is considered an advantage in terms of college admission or evidence of scholarliness. Additionally, the number of exams that a school administers, along with the number of AP 5s earned at a school, regularly appears in the media as evidence of quality for a school or school district and may even be used as a measure of the quality of the AP teacher. These are not quality indicators encouraged by the College Board but are definitely part of the landscape surrounding the AP programme and education in the United States. As a result, the standard-setting process is of critical importance. The scores' distributions that result from the standard setting final outcome are closely scrutinized by stakeholders, with any increases or decreases swiftly questioned. It is important to note that there is no optimal score

distribution or target. The standard-setting process is criterion-referenced, and if all students assessed meet the level of performance to receive a score of 5, then that is what would be reported. Despite the many proponents of AP, conflicting points of view do exist.

As previously reported, the AP programme recommends that a student may begin to receive college credit with a qualifying score of 3 on an AP exam. However, not all institutions agree. Some institutions are wary of credit by examination in general and accept very few, if any, AP scores or may only accept scores and reward general credit rather than credit in the subject area tested. At times there is a territorial issue or disbelief that anyone can prepare students as well as the course professor despite research showing students receiving AP credit are as successful in the next course as the students who took the introductory level course at the institution (Morgan and Klaric, 2007). At other times a prestige issue may factor into decisions about AP score acceptance, with institutions considering themselves more prestigious or above the norm only accepting scores of AP 4 or AP 5. At times these decisions are made in a vacuum with little or no data to support the decision; in some cases there may be a study that has called into question past student performance that is being cited, or decisions may be made in reference to experiences with previous processes like the former college comparability studies.

The AP programme values diversity and has as an on-going mission to attract more diverse groups of students into the AP courses. The volume of students taking an AP course and exam has continued to increase each year, and with that increase in students the number of diverse students has also increased. But there is still room for improvement. For certain AP courses part of the growth has been through an increase in younger students taking the courses during their 9th or 10th grade years in secondary school. For courses with large numbers of younger students, special analyses are conducted during the standard setting to compare the performance of what is considered the typical AP student to the performance of the younger students. It is very important that the rigour of AP be maintained and that a change in the population taking the course and exam does not cause a decrease in rigour. This is one of the key reasons using a criterion-referenced standard setting method is important and the reason that special analyses on student performance are conducted for the courses where the population has changed to include much younger examinees. On a similar but different issue, AP Language and Culture exams also conduct special analyses separating native speaker performance from non-native speakers to ensure

the exam is fair and a good measure of the construct and not inflated due to the inclusion of native speakers in the population being examined.

The AP Course and Exam Redesign and the move to standard setting have had the effect of causing educators to rethink their positions on AP score acceptance. As part of the standard setting, the SMEs must take the exam and then throughout the process they become very familiar with the ALDs, exam and level of rigour. This is an eye-opening experience and, for many, the first time they have ever seen an AP exam. It is not uncommon for the staunchest critic to have reversed their attitude in support of the AP course and exam by the time the standard setting concludes. It is still early in the process, but there is hope that as the true rigour of the AP courses and exams become more widely known, especially with recent changes after the redesign, more institutions will re-evaluate their policies to the benefit of deserving students who can be successful in the subsequent course, both potentially shortening their time to graduation and/or allowing the student the opportunity to use that gained freedom in programme of study to explore more advanced subject matter.

References

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (2014) *Standards for Educational and Psychological Testing*. Washington, DC: AERA

Angoff, W.H. (1971) 'Scales, norms and equivalent scores'. In Thorndike, R.L. (ed.) *Educational Measurement*. 2nd ed. Washington, DC: American Council of Education, 508–600.

Cizek, G.J. (ed.) (2012) *Setting Performance Standards: Foundations, methods, and innovations*. 2nd ed. New York: Routledge.

Hambleton, R.K., Pitoniak, M.J. and Copella, J.M. (2012) 'Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results'. In Cizek, G.J. (ed.) *Setting Performance Standards: Foundations, methods, and innovations*. New York: Routledge, 47–76.

Hendrickson, A., Ewing, M., Kaliski, P. and Huff, K. (2013) 'Evidence-centered design: Recommendations for implementation and practice'. *Journal of Applied Testing Technology*, 14, 1–27.

Huff, K., Steinberg, L. and Matts, T. (2010) 'The promises and challenges of implementing evidence-centered design in large-scale assessment'. *Applied Measurement in Education*, 23 (4), 310–24.

Kane, M. (2001) 'So much remains the same: Conceptions and status of validation in setting standards'. In Cizek, G.J. (ed.) *Setting Performance Standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, 53–88.

Kolen, M.J. and Brennan, R.L. (2004) *Test Equating, Scaling, and Linking: Methods and practices*. 2nd ed. New York: Springer.

Lacy, T. (2010) 'Examining AP: Access, rigor, and revenue in the history of the advanced placement program'. In Sadler, P.M., Sonnert, G., Tai, R.H. and Klopfenstein, K. (eds) *AP: A Critical Examination of the Advanced Placement Program*. Cambridge, MA: Harvard Education Press, 17–42.

Mislevy, R.J., Steinberg, L.S. and Almond, R.G. (2003) 'On the structure of educational assessments'. *Measurement: Interdisciplinary Research and Perspectives*, 1 (1), 3–62.

Morgan, R. and Klaric, J. (2007) *AP Students in College: An analysis of five-year academic careers*. College Board Report No. 2007–4. New York: The College Board. Online. https://files.eric.ed.gov/fulltext/ED561034.pdf (accessed 19 June 2018).

Morgan, D.L. and Perie, M. (2013) 'Setting standards in education: Choosing the best method for your assessment and population'. *International Journal of Science*, 3, 8–30. Online. https://issuu.com/ijosc.net/docs/www.ijosc.net (accessed 19 June 2018).

Pitoniak, M.J. and Morgan, D.L. (2012) 'Setting and validating cut scores for tests'. In Secolsky, C. and Denison, D.B. (eds) *Handbook on Measurement, Assessment, and Evaluation in Higher Education*. New York: Routledge, 343–66.

Pitoniak, M.J. and Morgan, D.L. (2017) 'Setting and validating cut scores for tests'. In Secolsky, C. and Denison, D.B. (eds) *Handbook on Measurement, Assessment, and Evaluation in Higher Education*. 2nd ed. New York: Routledge, 235–58.

Plake, B.S. and Cizek, G.J. (2012) 'Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods'. In Cizek, G.J. (ed.) *Setting Performance Standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, 181–200.

Context and Change in Standards Setting

Eva L. Baker

I read the contribution of Deanna L. Morgan with interest as the College Board and the Advanced Placement (AP) programme have venerable histories in the United States and continue to have generational impact (my two high school-aged grandchildren are taking seven AP courses between them this year). I will start with some particular comments about the essay, and then move to concerns associated with standard setting. I would place the standards-based focus much earlier than No Child Left Behind (2001). In particular, the National Council on Education Standards and Testing published a set of recommendations in 1992 (Raising Standards for American Education, 1992) that was the culmination of earlier work by Governors at the President's Educational Summit, 1989, followed by deliberations of the National Educational Goals Panel. The Report of the Standards Task Force (NCEST, E-1-19) provides the blueprint for Standards requirements that were developed for use in the Improving America's Schools Act of 1994, an earlier version of the Elementary and Secondary Act. Rereading this history can illustrate where the standards discourse came from and how practical implementation of standards in increasingly fractious contexts pushed a well-conceived idea off the rails. The quick lesson is that content standards were meant to fit into a broader set of expectations, for schools, instruction and equity, in addition to school subjects. Nonetheless, discord allows the College Board to rightly claim a national, voluntary curriculum, at least for a particular segment of students.

Setting standards

The document clearly reports the nature of the AP assessments and their uses. It describes both the reasons and the detailed process of setting standards. The discussion of the shift from comparability of college performers and the more recent panel approach to standard setting is plausible. However, any criterion that uses college grades will embed the same issues of variation among professors' judgements. For instance, if grade inflation is a fact, it probably needs nonetheless to be included as part of the process.

The use of a panel of subject matter experts (SMEs) may avoid some of the difficulty of college comparability, but it raises new and difficult issues.

One of these is that the rules for categorizing performance (grades) is less transparent than before, as it is no longer grounded in real performance, but in estimates of types of student performance by the SMEs. The availability of the achievement level descriptors (ALDs) seeks to provide a common understanding. Yet, in some tests, these descriptors are generated post hoc by reviewing items, while in others, the ALDs are part of the design process. Although classified as criterion-referenced, the estimation process has a normative component, that is, what proportion of borderline students would get the item right.

The procedures of Rounds 1 and 2 are carefully described. However, it falls to the AP Program staff to design cut scores, and illustrative information that goes into these decisions would be helpful. The description of AP Capstone performance tasks is provocative, especially the adjustment of profile assignments based on expected distributions.

The saving grace of the AP programme and its examinations is that it has curricular relevance and exams are not free-standing. Further exploration of validity issues would be a welcome focus for the future.

Reference

NCEST (National Council on Education Standards and Testing) (1992) *Raising Standards for American Education*. Washington, DC: National Council on Education Standards and Testing. Online. https://files.eric.ed.gov/fulltext/ED338721.pdf (accessed 19 June 2018).

Filling the aligned instructional system void: AP courses and exams in US high schools

Betsy Brown Ruzzi

In her chapter, 'Setting standards in the United States: The Advanced Placement programme', Deanna L. Morgan describes how the Advanced Placement (AP) courses and examinations play an important role in preparing US students for selective universities. These courses and examinations are provided, for a fee, by the College Board, a not-for-profit organization based in New York City, to US high schools that choose to offer as few as one or up 37 courses to their high school students in Grades 9 through 12. Students who take an AP course can decide to take the examination or not, if their high school does not mandate test taking. In some cases, the school district will pay for low-income students' examination fees. But in most cases, the student or parent pays that fee which is $94 per test. It is not unusual for US high school students to take as many as ten AP courses and their accompanying exams as one way to set themselves apart in the race to be admitted to a highly selective university.

Why AP?

Unlike top-performing education systems around the world, the United States does not have a common programme of study that all students experience during compulsory school. Most top-performing education systems have built aligned instructional systems made up of common courses with accompanying syllabi, curriculum frameworks matched to the syllabi, assessments that measure what is taught in the curriculum and examples of students' work, with commentary, that show what it means to succeed on the assessments. The US state of Massachusetts comes closest to having an aligned instructional system, and the state's performance on the US National Assessment of Education Progress (NAEP) demonstrates the benefit of that near alignment, putting the state at the top of US student performance. However, the state's course coverage and assessment design are yet to match top-performing systems. As a consequence of having

no common programme of study, no common curriculum, or common assessments that can, using a common metric, show what a US student knows and can do, the AP programme has filled a void for US colleges and university admission's programmes. As stand-alone, curriculum-based examinations of 'college level' work, the AP courses and exams are not a diploma programme or a qualification. Instead, AP provides high school students, when available, with common course syllabi and examinations in 37 subjects. In their own way, APs serve as the nation's high school leaving examinations for students wanting to show how they measure against other high-achieving students and demonstrate their college readiness.

How are APs used in the US?

AP examinations are used in a number of ways by students, high schools, colleges and universities. University admissions offices use AP course completion and exam scores as one sorting mechanism in their selection of candidates for admission. They also use exam scores as a course placement tool for first year students. Success on AP exams is one way for students to earn college credit prior to entering university that, in some cases, either shortens university, and therefore the cost incurred to students, or allows students to take higher level college courses upon entry. And US high schools offer AP courses and exams to provide a challenging pathway for high school students who are ready for accelerated learning.

Are there other benefits to students who take AP courses and exams?

In addition to helping students in the college admissions process and providing college credit where available, other benefits gained from taking AP courses in US high schools are: the exposure to challenging materials; the high expectations for what students can do by teachers; and the participation in classes with highly capable peers. Much like assessments used in top-performing education systems, AP examinations are common across the country, scored using a common rubric and awarded on merit. Lessons from the AP Program can certainly be extended down into the early grades if the US wants to implement one of the elements found in top performing education systems – highly aligned instructional systems.

Part Three

Differing measures
and meanings

3

Chapter 14

The meaning of national examination standards

Jo-Anne Baird

We *all* know what examination standards mean

Various understandings of what examination standards mean have been evident in the research literature and the public discourse for some time. At times these definitions could comfortably coexist, but oftentimes they are fundamentally incompatible. Let us consider a few ways in which they are discussed. Do examination results represent intelligence – a stable feature of students that is assumed not to change much over different cohorts taking the examinations? Are examination results an indicator of attainment, which can improve (or decline) between years depending upon factors such as quality of teaching and student motivation? When results have risen, have examination boards made the examinations easier, perhaps for commercial or political advantage? Would it be feasible for all examination candidates to pass, so long as they met the criteria? Are examination grades essentially a quota for university entrance? Do examination grades ensure progression standards for universities? Each of these ways of thinking about examination standards has implications for policy and practice, as well as theoretical implications. The very fact that different uses of the term 'examination standards' coexist and can be compatible or incompatible needs some explanation.

With high expectations for the knowledge economy, education systems and their examinations can come under considerable criticism. In the project reported in this book, many examination boards were under a range of pressures from stakeholders stemming from their dissatisfaction with examination standards. Political pressures were being felt in a number of countries, either to change the examination structures or the outcome standards. Grade inflation was not an issue for those systems that maintain similar proportions of students gaining the grades each year, but in other cases it had been a major focus of debate. Examination boards were also anticipating where future challenges were likely to arise and preparing to address them, such as what evidence could be marshalled to show that students with the same grades in different years' examinations had similar

performances. Not only did participants in the project have different definitions of examination standards and a variety of pressures from stakeholders, they had a range of ideas about what constituted rigour in setting standards. Participants' beliefs regarding rigour no doubt stemmed from the paradigm in which they were most comfortable operating and its attendant methods, as discussed in Chapter 1.

Meritocracy and social mobility operate on the notion that students' qualities and efforts will be recognized through a fair system, and examinations act as a tool for a fair system in many societies. Given the onus upon the examinations to deal out life chances fairly, how they are defined and set is hugely political, even if the debates are at times muted. In this chapter, the evolving literature on definitions of examination standards is traced. Over time, the literature can be characterized as the rise of psychometrics, outcomes-based and latterly curriculum-based and psychometric systemic definitions. The meaning of examination standards is essential for comparative purposes; we want to know that the grades mean the same thing across different students, versions of a public examination and so on. Thus, we next turn to the meaning of comparability of standards and show that setting lofty theoretical ideals for what counts as comparable is ultimately unhelpful for exam boards which need to deliver comparable standards under real-world conditions that do not meet these strictures. An ecological model of examination standards is outlined, which serves to organize the literature and explain why different definitions of exam standards coexist and why examination boards and other stakeholders often draw upon a range of definitions, even if they do not always recognize this. Examination boards are responsible for standards at all levels of the ecological model and therefore have to be able to defend them at each level. Definitions are associated with the paradigms set out in the introductory chapter. Finally, we classify the methods used by some of the countries involved in the project.

In Chapter 1, we outlined three assessment paradigms and recap them very briefly here as a reminder. The first was the *psychometrics* paradigm, arising from psychological theory and methods and relying on particular statistical techniques, applied to groups of students. The second was the *outcomes-based* paradigm, which has its roots in Taylorism and vocational assessment. Outcomes-based assessment depends upon qualitative judgements of experts and can most easily be applied to small numbers of students, often using observational methods. The third was the *curriculum-based* paradigm, which arose in education, is typically applied to large-scale assessments such as public examinations but can be used by teachers with smaller groups. Statistical techniques are also central to this method.

Definitions of examination standards

Most of the research literature on examination standards has been written from the psychometric paradigm (e.g. Blömeke and Gustafsson, 2017; Cizek and Bunch, 2007; Hambleton and Pitoniak, 2006). Lawn's (2008) analysis of the Americanizing of the field is pertinent. Scientific publications are generally dominated by the US due to the use of English as the *lingua franca* of science, the economic dominance of the US and its massive population. Chapter 4 described the kinds of techniques developed within the psychometrics paradigm. Encountering the research literature, the reader could be forgiven for thinking that there was, in fact, only one way of thinking about examination standards. A view often encountered is that the psychometrics paradigm is more scientifically advanced and that all other approaches are inferior. We return to these points later. Research publications are voluminous and the literature is fragmented, so it is difficult for researchers, let alone practitioners, to get an overview of the field. Most people are working in a context that does not expose the paradigm that they are working in because the historically and culturally bound systems largely serve to maintain the status quo (see Chapter 15).

On discovering that others approach assessment differently, a typical response is to express surprise, be quizzical and to propose the adoption of our own tried and tested model. These encounters can sometimes become heated because they challenge beliefs about our professional understanding and therefore our identities and status. Having built a career around particular knowledge, technical processes and ways of thinking, it is unpleasant to say the least to have someone claim that it is incorrect, or even irrelevant.

An example from personal experience was a project on the development of national basic numeracy tests in the late 1990s in England. All parties in the project struggled to work together. The project team involved policymakers who were steeped in an outcomes-based approach, was led by researchers from a curriculum-based perspective and as a member of the research team, I was essentially arguing for a psychometrics paradigm. These perspectives influenced everything from item design, development of the tests and standard setting to how the test results were interpreted. From the examination board representative's perspective (curriculum-based), a big problem and a risk for the project was poor test design. From an outcomes-based perspective, test development could be conducted by hiring appropriately qualified individuals to write the tests, in keeping with pre-specified criteria for functioning numeracy. Questions would be set in real

life contexts in the test, such as interpreting a train timetable or handling money. The test would then be sat by the intended cohort and the pass mark could be specified in advance in accordance with the criteria. With a mastery approach, the pass mark was likely to be set in advance of 80 per cent to ensure that candidates had sufficient knowledge of the required subject matter to be able to handle numeracy in everyday life. Now, from a curriculum-based or psychometrics perspective the meaning of the resulting standards would be questionable. Current thinking was that the difficulty of examinations was often not as intended by the test-writer and therefore standard setting processes are needed to adjust for this between different sittings. Expectations around public examinations required the standards to be similar, 'fair', between sittings. Therefore, the outcomes-based perspective was far from compatible with either the curriculum-based or psychometrics perspectives. Wolf and Cumming (2000) described similar situations in England and Australia in the development of basic skills tests. Uncomfortable compromises need to be made to produce examinations that everyone can live with when working across paradigms. This situation persists for basic tests of literacy and numeracy in England some 20 years on.

Comparisons across time and cultures can reveal paradigms. In England, a body of work has been published on the meaning of examination standards from an examining tradition (Baird and Gray, 2016; Baird *et al.*, 2000; Baird, 2007; Cresswell, 1996; Christie and Forrest, 1981; Coe, 1999, 2007, 2010; Newton, 1997a, 1997b, 2003, 2005, 2010a). As discussed in Chapter 1, the examining tradition comes from an education discipline perspective and has tended to view standards as properties of the cohort of students taking the examinations rather than a matter solely regarding individual students. England has experienced two paradigm wars in examination standards. The first, in the early 1980s, related to the use of psychometrics (Panayides *et al.*, 2010), specifically the Rasch model. Arguments against the use of psychometrics were statistical as well as educational. At that time, psychometric techniques only operated with multiple choice tests, and there were concerns that this would have a damaging effect upon education, with the curriculum being fragmented and students being taught to the test. Educationalists won the argument, but the curriculum-based paradigm faced a subsequent challenge. The outcomes-based assessment paradigm swept the globe, through its notion of criterion-referencing and competency assessment, during the 1980s and 1990s. A culmination of the war between the curriculum-based paradigm and the outcomes-based assessment paradigm was the internal organizational strife for public examination boards in England and Scotland in the early part of

this century when they were required to merge with a vocational awarding body. Evidence for the rejection of a pure form of outcomes-based standard setting techniques (criterion-referencing) featured in the literature at that time (e.g. Cresswell, 1987, 1994; Wolf, 1995). Outcomes-based approaches were adopted in Scotland, South Africa and New Zealand (see Chapter 15). Let us turn to the definitions of examination standards previously published and the progress that has been made in the literature, since there have been some key developments.

Comparability of examination standards

As discussed in Chapter 1, much of the pressure to define standards and to set them at a particular level comes from the need for comparability. In the psychometrics literature, this is termed invariance. Due to the uses of examination scores or grades for entry into higher education, by employers, for school funding and accountability and so on, comparability of outcomes between years, across subjects, between qualifications and so on, are variously required in different systems. The emphasis given to each form of comparability varies culturally, including over time. The term 'invariance' refers to the notion that examination scores should mean the same, should not vary, over these conditions.

In the psychometrics tradition, ensuring that the standards are the same between tests is termed 'equating'. When equating is conducted, it is assumed that the underlying attributes (the construct) are the same in the two tests. Further, the difficulty levels of the two tests should be approximately the same to begin with, the same populations should have taken the tests and the reliability of the two tests should be the same. As described in Chapter 4, a range of techniques can be deployed for test equating. But there are broader requirements on comparability of national examination outcomes than equating can contend with. When students apply for a university place, they may come with different subject results. How should the university treat applicants with an economics grade rather than a geography grade? Holland (2007; Table 14.1) termed the techniques to deal with these broader issues scale aligning, indicating the type of linkage that had been developed for each situation of variation in constructs, difficulty, test-taking population or reliability of the tests. At the bottom of Holland's table (Table 14.1) there is a 'prediction' category in which all of the assumptions are open. This is a continuum model of comparability in which not all forms are created equal; equating is the purest form with the highest level of assumptions met regarding the similarity of the two sets of test results being compared (Newton, 2010b).

Table 14.1: Holland's (2007) equating quality continuum

Quality of Link	Linking Category	Constructs	Difficulty	Population	Reliability	Type of Link
Highest	Equating	same (intended)	same (intended)	same (intended)	same (intended)	equate
	Scale aligning	similar	similar		similar	concordance[i]
		same (intended)	similar	same (intended)	different	calibration[ii]
		similar	different	different	similar	vertical scaling
		different	–	common	–	battery scaling[iii]
		different	–	different	–	anchor scaling[iv]
Lowest	Predicting	–	–	–	–	prediction

[i] Concordance is most similar to equating but it is recognized that since the populations taking the tests are different in nature, the equating link is not as strong.

[ii] Calibration is the term used when the reliability of the tests being linked is known to differ. Vertical scaling is the term used when the difficulty of the tests differs as they are designed for different populations

[iii] Battery scaling is the term used when the constructs on tests differs. In battery scaling the same population has taken the tests but there are no assumptions regarding difficulty or reliability

[iv] Anchor scaling is the term used when the constructs are different and different populations take the tests. No assumptions are made regarding common levels of difficulty or reliability.

Many of the ways in which examination results are used around the world require invariance, or comparability in circumstances that do not meet the strictures of equating. Examination boards are grappling with the meaning of examination standards in conditions that look more like the bottom of Table 14.1. What good then is a deficit view of the real world problems that examination boards face? This approach is a technically purist, but ultimately partial worldview for setting standards in national examinations. One approach to this conundrum is to take the position that assessments should be created so that they meet the strictures of the psychometrics paradigm (Andrich, 2004). In other words, the instruments can be constructed to fit the model which has invariance built into it. Self-evidently, though, the kinds of invariance that we seek go beyond comparing

the same or even similar constructs, so to take this position, the notion of what is tolerably similar has to be stretched. Alongside the psychometrics tradition, a separate literature developed in England over the past 20 years, which sought to grapple with meanings of examination standards.

Levels of description

Before explaining the definitions that have been proposed in the literature, we first have to explain the current state of this field and to introduce a new way of organizing the proposed definitions. Examination boards looking to the research literature to find a ready definition of examination standards would find great difficulty. In signing up to a specific definition, such as criterion-referencing, it would appear that an examination board cannot predict or control the expected distribution of student results. Equally, norm-referencing has little to say about students' performances. Since challenges to examination standards can come in a wide variety of guises, this is problematical. Examinations play a social function and examination boards answer to a wide range of stakeholders. As such, examination standards need to be understood within the societal context of their operation. Gone are the days when examinations were only about the students taking the tests. Therefore, to understand examination standards, we need an ecological model.

Ecological models provide a framework for understanding the various, interacting systems that determine individual lives. By including the various layers of human society, ecological models consider contextual or environmental factors in the analysis of individuals. At the same time, ecological models do not claim that the multiple levels or systems are static and unchanging. Rather, these models acknowledge that systems in turn interact and influence each other (Bronfenbrenner, 1974).

One of the most widely applied models in the social sciences is Bronfenbrenner's (1994) socio-ecological model of child development, based on ecological systems theory. In this model 'the ecological environment is conceived as a set of nested structures, each inside the other like a set of Russian dolls' (Bronfenbrenner, 1994: 39). At the centre of Bronfenbrenner's model is the micro-system, which takes into account individual characteristics such as age, gender and language. The meso-system is the immediate social and physical environment of the individual and includes for example family, friends and school. At the exo-system level are structures and institutions of society that govern the meso-system. This would include government agencies or the distribution of goods and services. The fourth level is the macro-system, which encompasses the

structures and ideologies of overarching institutions such as the economy or the educational system (based on Bronfenbrenner, 1976: 5–6). A fifth system called the chrono-system is the level of changes or continuities in individual and community life over time, such as resettlement or changes in socio-economic status.

Various social sciences have developed and adapted ecological models for decades. Even before Bronfenbrenner applied his socio-ecological model to childhood development (1974) and education (1976), it was used by social psychologist Kurt Lewin in the 1940s and 1950s. Ecological models have also been used by other psychologists (e.g. Barker, 1968; Gibson, 1979), social psychologists (e.g. Pappaport, 1987) and their use is advocated in social work theory (e.g. Payne, 2014). Additionally, ecological models have found regular application in social and public policy research, particularly in areas of public health (e.g. McLeroy *et al.*, 1988; Stokolos, 1996) and in education (e.g. Hodgson and Spours, 2015). Building on earlier research that uses ecological models in language assessment (McNamara, 1997, 2007), Zumbo *et al.* (2015) developed a model of differential item functioning (DIF) based on an ecological model. They suggest that there are five levels that explain the difficulty of a test item: '(a) test format, item content, and psychometric dimensionality; (b) person characteristics and typical individual differences variables such as cognition; (c) teacher, classroom, and school context; (d) the family and ecology outside of the school; (e) characteristics of the community, neighbourhood, state, and nation' (Zumbo *et al.*, 2015: 140). With this model they move beyond traditional literature on DIF that focuses on the innermost level. They maintain the view 'that neither the test taker nor the cognitive processes in item responding are isolated in a vacuum' but that test-takers approach each item based on their 'social and cultural present and history' (ibid.: 140).

At the centre of an ecological model on educational assessment is the examinee (Figure 14.1). Mostly when we think about educational assessment, discussions centre on this level. Issues related to student anxiety, performance and so on are at the examinee level. However, student experience is embedded within particular examination systems in different countries, each of which has its own way of operating. As discussed above, each high stakes, large-scale, school leaving examination system has to articulate with the education system in which it operates so that the grades are useful to stakeholders. Features of the education system, its structures and processes, will affect the suitability of different examination systems. Education systems themselves are defined by the wider social and cultural

contexts that they inhabit; thus oral traditions of education and assessment are more suitable in some societies than in others. Finally, societies change over time, so historical practices may well impact upon assessment systems, but we can also anticipate evolution, even if it is slow in many cases (see Chapter 15). In the next section, various definitions of examination standards are outlined, beginning with those classified at the level of the examinee. None of the definitions are classified at the education system level, but that is retained in the ecological model because it constrains the shape and function of examination systems.

Proposed definitions may be complementary at different levels in an ecological model of examination standards. We are surely interested in what examination standards have to say about individual students, but also about the cohort as a whole and, more widely, how the standards articulate with the education system and the wider context of the examinations. However, definitions across levels may also be in tension or contradictory. Most previously proposed examination standards definitions can be classified as being within a particular level, but some of them speak to standards across levels. In the sections following, we discuss the definitions at each level, but first we outline issues in the state of the field generally.

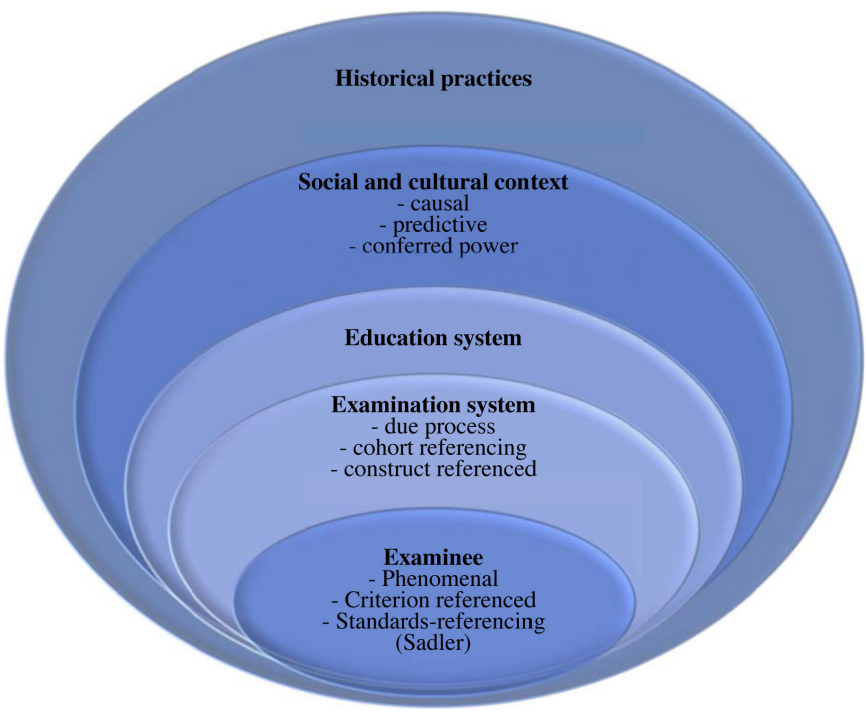


Figure 14.1: Ecological model of examination standards definition

As is common in many social science fields, rather unhelpfully, authors have used different terms for the same, or very similar, definitions of examination standards. The terms *social comparability* (Cresswell, 1996), *conferred power* (Baird *et al*., 2000) and *conventional comparability* (Coe, 2010) and are closely linked for example (see Table 14.4). Alternatively, authors have used the same term – standards-referenced – to mean different things, as discussed below.

Another problem is that sometimes the method of setting standards was confused with the definition of examination standards (Newton, 2010a). For example, *statistical comparability* (Coe, 2010) is a method rather than a definition per se. Newton's (2010a) definitions were designed to supervene methods and to address the underlying meaning of the standard. For example, his phenomenal definition indicates features of the candidates' attainments, but those features may be gleaned by a range of methods such as qualitative judgements of performances, statistical methods or otherwise. Conceptually, methods and definitions are distinct, but it is obvious that some methods align better with certain definitions. For example, it is a stretch to envisage a statistical approach to criterion-referenced assessments. Other definitions, such as the *conferred power* definition (Baird, 2007), could be enacted equally by subject matter experts or statisticians using qualitative or statistical methods because the definition simply relies upon particular individuals being deemed to have the power to call the shots. Lack of an agreed classification scheme for the meanings of examination standards is troublesome for the research field and practitioners alike and is indicative of a field that is struggling to codify its terms and has multiple paradigms at play.

Examinee definitions

At the level of the individual examinee, three definitions for examination standards have been proposed (Table 14.2). Criterion-referencing has long been discussed as an assessment approach and has been proposed as a definition of standards (e.g. Wiliam, 1996). This approach became very attractive as a response to statistical methods used in isolation. People worried that even if the same proportion of examinees were being given the grades each year, their skills were deteriorating; they could not spell or were not as numerate. In the criterion-referenced approach, a written description of the skills, competencies and understanding in certain areas of knowledge is used as a basis for judgements on the grade-worthiness of a student's performances. Used alone, criterion-referencing does not tackle the fact that examination questions vary in difficulty even when examiners

try to set equally difficult examinations. If invariance of demand of the examination is important, which it surely is in high stakes school leaving examinations, criterion-referencing is problematical because the standards vary between examinations (Cresswell, 1996). To tackle this issue, criterion-referencing was subsumed into a wider approach, first dubbed weak criterion-referencing (Baird *et al*., 2000) and later termed attainment-referencing (Newton, 2011).

Standards-referencing (Sadler, 1987) was distinguished from criterion-referencing in that criteria for several grades were elucidated rather than simply pass/fail, there was a recognition of tacit judgements, a series of judgements were required rather than judgements of one-off performances, and exemplar performances were utilized as part of the process of developing an understanding of the criteria. This examinee-level definition of standards, which is well articulated in the higher education assessment literature, was appropriated and its meaning extended in the psychometrics tradition as opposed to how Sadler (1987) used the term (see Table 14.5).

Table 14.2: Examinee level definitions of examination standards

Term (paradigm)	Definition	Similarly graded students share similar	Example source text	Similar to
Criterion-referenced (outcomes-based)	Performance meets the pre-determined criteria	Performances in relation to written criteria	Popham & Husek (1969)	Domain referenced (Christie & Forrest, 1981) Standards referenced (Sadler, 1987) Performance comparability (Coe, 2010)
Standards-referencing (outcomes-based)	Configuration (or pattern) of performances over a series of testing episodes and tasks	General performances in relation to a specified level of criteria	Sadler (1987)	Limen referenced (Christie & Forrest, 1981) Criterion-referenced
Phenomenal (n/a)	Features, properties or dispositions that comprise attainment	Similar learning outcomes	Newton (2010)	All examinee level definitions

In both moves – from criterion- to attainment-referencing and standards-referencing appropriation – there is a shift from an examinee-level definition of examination standards to systemic, integrated definitions. Further, there is an annexation of outcomes-based paradigm techniques within curriculum-based and psychometrics paradigms. We return to this issue later.

Newton's (2010a) phenomenal definition is an umbrella term for all definitions that make claims for examinee-level properties of attainment. The phenomenal (Table 14.2), causal and predictive (Table 14.4) definitions proposed by Newton (2010a) were designed to be independent of standard setting methods and therefore are not necessarily aligned with a specific paradigm. Criterion- and standards-referencing as defined in Table 14.2 are outcomes-based paradigm definitions.

Examination system definitions

Not all definitions of examination standards are at the level of the examinee (Table 14.3). A clear example is Cizek's (1993) due process definition in which grading is not open to challenge so long as the agreed procedures have been followed. This is a legalistic definition, generated in a US litigious context, but it has utility in almost all assessment settings and is sometimes resorted to at times of challenge. However, as the due process definition does not in itself spell out what the content of the standards is, it is itself open to challenge in a wide range of ways: the outcomes, content standards and the processes themselves. With declining reverence for authority in many societies, it is less likely that a position which relies upon pointing to the procedures alone will be acceptable. As this definition is about procedures, it says little about standards at the examinee level. Instead, it specifies standards at a systems level. Also, it can be allied to any of the three paradigms, as due process could be generated from the outcomes-based, curriculum-based or psychometrics paradigms.

In cohort-referencing, the same proportion of candidates is awarded the grades at each sitting, no matter their performances or the nature of the group sitting the examination. Historically, for A level examinations in the UK, there was a policy of awarding the top 10 per cent of candidates a grade A, the next 15 per cent a grade B and so on (Christie and Forrest, 1981: 13; this was only ever a 'rough indication' and not a strict policy). The term 'norm-referencing' is often applied to this approach, but as Wiliam (1996) pointed out, norm-referencing involves using a random sample of test-takers to establish population norms and using those as standards for this year's test-takers using the same test. Therefore, cohort-referencing is a more accurate description of the method that is used for

school leaving examinations. Introduction of criterion-referencing and standards-referencing was largely in reaction to this approach due to its lack of consideration for candidates' performances. After all, teaching, motivation and even school attendance could wither under this definition but the grading would remain the same. Cohort-referencing sits within the curriculum-based paradigm while norm-referencing is within a psychometrics paradigm. We have not found an example of norm-referencing being used in public examination standard setting, but as we will see later, there are a number of examples of cohort-referencing.

Table 14.3: Examination system level definitions of examination standards

Term (paradigm)	**Definition**	**Similarly graded students share similar**	**Example source text**	**Similar to**
Due process (n/a)	Grades are issued according to pre-codified rules and procedures	Specified by the process	Cizek (1993)	–
Cohort referencing (curriculum based)	Proportion of candidates awarded each grade remains the same	Standing (within population taking the examination)	Cohort referencing (Wiliam, 1996)	Norm referencing (Christie & Forrest, 1981) No-nonsense, equal attainment (Cresswell, 1996)
Construct referenced (psychometrics)	Same underlying ability (statistically, psychometrically), taking into account difficulty of the items	Levels of a latent trait	Wiliam (1996)	–

Construct-referencing is essential to the psychometrics paradigm. In this approach to standards, examinees are worthy of the same grade if they have the same level of the underlying trait that the examination is assessing, such as ability in music. Psychometric statistical models assume that a student's ability is defined by their scores on the questions and, at the same time, that the difficulty of the questions is defined by students' success rates. Therefore, a student's latent ability level is defined in relation to others who took the questions. This makes a construct-

referenced approach an examination system level definition because students' abilities (the standards) are defined within a frame of reference that includes the characteristics of the test-takers and the items, as well as the interaction between them (Andrich, 2018). All statistical models make assumptions, and there have been debates about the extent to which these can be sustained or are helpful in setting examination standards (e.g. Goldstein, 1979; Goldstein and Wood, 1989).

Social and cultural context definitions

Yet other kinds of definitions go beyond the examination system, recognizing that there are features of groups of candidates that cause examination performances (Table 14.4). These definitions could have been classified at the education system level, as their proponents often point to educational causes of examination performances. In the catch-all definition, examinations are deemed to have comparable standards if they have the same distribution of grades having taken into account characteristics of the candidates taking the examinations. The causal definition is similar. Other definitions (e.g. examinees having attended similar schools) are essentially partial attempts to capture some of the causal variables in a pragmatic way. This immediately begs the question of which causes should be taken into account. Newton (2010b) argued that only direct causes should be taken into account, but in turn we can question what counts as a direct cause, since a wide variety of variables, such as emotional state, have been found empirically to affect examination outcomes (Baird, 2010). Therefore, these definitions are classified here at the social and cultural context level of the ecological model because the features that could be taken into account are open to dispute and may well take in socio-economic status or other factors that are not delimited by the education systems. Further, deciding which variables should be taken into account in these models is likely to be culture-sensitive.

Some variables will be politically acceptable in some contexts and not in others. This is most easy to see in terms of historical practices, but we do not classify these definitions at that level of the ecological model because our focus is upon current definitions in use. To illustrate a problematical variable, though, let us take gender. Allocation to secondary school type (grammar or secondary modern) was determined by outcomes of the 11+ examinations in England in the 1970s. Routinely, the pass mark for boys was lower than that for girls because it was believed that although girls did better than boys on average in the examinations, their ultimate performances in the education system and the world of work would not reach that of their

male counterparts. Explanation for these effects was biological rather than social; it was thought that males were late developers and should not be constrained at age 11 by their performances on the examination. Taking a sociocultural lens to this yields a rather different interpretation of the examination results and the use of the gender variable to control them, and these practices were dropped. In the catch-all and causal models, so long as a variable can be shown to affect outcomes empirically, it is fair game for inclusion in the model. However, a more theoretically driven approach to model-building is called for because a wide range of variables are associated with examination performances that have questionable validity in these standards definitions (e.g. students' mood, comfort of clothing). The causal approach is based upon the curriculum-based tradition.

Table 14.4: Social and cultural context definitions of examination standards

Term (paradigm)	Definition	Similarly graded students share similar	Example source text	Similar to
Causal (curriculum based)	Groups of students with the same characteristics are awarded the same grades on average	distributions of ability and prior attainment, attended similar schools with identical entry policies, were taught by equally competent teachers and were equally motivated (as a group)	Cresswell (1996); catch-all	Causal (Newton, 2010) Statistical comparability (Coe, 2010) Same-candidates, Value-added, Similar schools (Cresswell, 1996)
Predictive (n/a)	Potential that is implicit in attainment	likelihood of future success	Newton (2010)	–
Conferred power (n/a)	Selected individuals are endowed with the responsibility for deciding the grade-worthiness of performances on the basis of their values	performances, as valued by empowered judges	Cresswell (1996); social	Social facts (Wiliam, 1996) Sociological (Baird *et al.*, 2000) Conventional comparability (Coe, 2010)

There are two further definitions proposed which are at the social and cultural context level of the ecological model that can be utilized by different

paradigms. The conferred power definition was discussed earlier. Newton's predictive definition involves standards related to how students perform in the future, such as in educational outcomes or in salary terms. This definition goes beyond educational assessment paradigms and deals purely with the consequences of standards. Although this definition is raised in the assessment research literature and in challenges to examination standards, no application of this definition has been encountered in this project in which it is used in practice to set standards for school leaving examinations, though it is often referred to when critiquing standards.

Systemic definitions of examination standards

Faced with all of these different definitions, what is an examination board to do? One way of thinking about what examination boards actually do in practice is to consider whether they are willing to disregard any of the definitions or juggle with multiple definitions. Going back to our ecological model (Figure 14.1), to what extent would an examination board be prepared to disregard challenges to the standards arising from different levels of the ecological model? For example, an examination board would surely feel obliged to have a response to challenges regarding the level of students' knowledge and skills. Likewise, if the proportion of students gaining the grades changed dramatically between years, it is reasonable to propose that in most countries, an explanation would be expected from the examination board. Additionally, stakeholders are likely to see themselves as entitled to transparent procedures being followed openly, and therefore examination boards are likely to defend their standards against due process challenges. They might also feel entitled to ask whether the people setting the standards were the right ones (conferred power), whether the nature of the group of candidates taking the examination had properly been accounted for (causal) or if the standards tell us anything about whether the students could cope with higher education (predictive), for example. Therefore, examination boards have to field challenges based upon a range of definitions. There are instances where initial responses to challenges have been rather narrow, and these have resulted in the demise of the senior managers. For example, when New Zealand introduced a criterion-referenced examination in 2003 with surprising results, the ensuing inquiry found that they had disregarded a causal definition of examination standards (see Chapter 15). Expectations for examination standards are very broad in stakeholders' minds, and examination boards have to balance these sometimes competing definitions in practice. This is a technical *and* political task.

Fortunately, two rather broad approaches have been proposed in the literature (Table 14.5). Both arise from the need to account for examinee performances and the difficulty of the examination. The attainment-referenced approach is derived from the curriculum-based paradigm and the standards-referenced approach comes from the psychometrics paradigm. They are very similar, though the standards-referenced approach need not be curriculum-related. Conceptually, the standards-referenced definition does not need to be based upon psychometric methods but the term is used in the psychometrics literature. This use of the term 'standards-referencing' is very much at odds with the origins in Sadler (1987), as he makes no mention of statistical methods at all. Attainment-referencing and standards referencing are mixed methods definitions (see Chapter 4), taking into account qualitative judgements and statistical information. These systemic definitions are also multi-level in terms of the ecological model because they incorporate examinee, examination systems and social and cultural context level definitions.

Attainment-referencing involves qualitative judgements of students' performances (examinee-level) and statistical information about the group of candidates taking the examination (examination system level; social and cultural context level). Standards-referencing involves qualitative depictions of students' attainment that require a qualitative approach to be incorporated in the standard setting system (examinee level) alongside a constructs approach (examination system). Thus, attainment-referencing and standards-referencing are mixed methods approaches (see Chapter 4).

Table 14.5: Systemic definitions of examination standards standards

Term (paradigm)	**Definition**	**Similarly graded students share similar**	**Example source text**	**Similar to**
Attainment-referencing (curriculum based)	Overall level of attainment in the curriculum being examined	Underlying attainment	Newton (2011)	Weak criterion-referencing (Baird *et al*., 2000) Standards-referencing (Cizek *et al*., 2004)
Standards-referencing (psychometrics)	Defined categories of performance	Latent trait levels	Cizek *et al*. (2004)	Attainment-referencing Weak criterion-referencing (Baird *et al*., 2000)

The focus of this book is on national examination standards, but there are of course international assessments such as those operated by the Organisation for Economic Co-operation and Development (OECD) or the International Association for the Evaluation of Educational Achievement (IEA). These organizations operate international large-scale assessments such as the Programme for International Student Assessment (PISA) or the Progress in International Reading Literacy Study (PIRLS), respectively. Both organizations use psychometric, standards-referenced approaches to the definition of standards. However, they take different approaches when it comes to situating the standards in context. For PISA, OECD claims that the standards are curriculum-unrelated and therefore can be applied unproblematically across jurisdictions. In PIRLS, recognition of the curriculum and context of the standards in different jurisdictions plays a much larger part of the test development process and the content of the performance standards. In both cases, the 'system' is international rather than at national level and, in effect, different positions are taken on whether the social and cultural context needs to be integrated into the content standards for the test to justify the outcome standards.

Examination board definitions in practice

In the case study chapters in this book, the authors have outlined the approaches to setting standards. These are collated in Table 14.6 and we have included other jurisdictions involved in the project. Several countries adopt a criterion-referenced approach, either in the national examinations or in their teacher assessments used for school leaving results. Cohort-referencing is also the preferred approach in a range of settings. Construct-referencing is the underpinning rationale for standards in the US Advanced Placement examinations and also for tests in Queensland and Sweden, used alongside teacher assessments. From previous work we know that Scotland (Baird and Gray, 2016) has adopted attainment-referencing as a definition of examination standards, in addition to its use in England. In Hong Kong, a standards-referenced approach is used. Note that although countries might share the same definition of examination standards, the methods they utilize to set standards may be very different, as in the cases of England and Scotland (Baird and Gray, 2016). France is included in the table as an outcomes-based approach, although Gauthier (Chapter 7) stated that there is no standard setting process for the baccalauréat, that it is only marked. As the points used are meaningful in terms of passing the baccalauréat, it is a criterion-referenced definition, even if questions remain about the approach to deriving and agreeing the criteria.

Table 14.6: Definitions of exam standards in different jurisdictions

Definition (ecological model level)	**Jurisdictions**	**Paradigm best fit**
Criterion-referenced (examinee level definition)	France Sweden (teacher assessments) Queensland (teacher assessments)	Outcomes based
Cohort-referenced (exam system level definition)	Chile Georgia South Korea South Africa Victoria	Curriculum based
Construct-referenced (exam system level definition)	Queensland (tests) US (Advanced Placement tests) Sweden (national tests)	Psychometrics
Attainment-referenced (systemic definition)	England Ireland Scotland	Curriculum based
Standards-referenced (systemic definition)	Hong Kong	Psychometrics

Conclusion

In this chapter we have investigated the different meanings previously published in the literature on examination standards. These have largely been derived from the US and English literature, though not always. Previously, there have been ill-fated attempts to categorize the different ways of defining examination standards, with each new article proposing a different system. Here, we rationalize the literature by showing that the definitions are associated with different levels of education and examining systems. Some define standards with regard to the characteristics of individuals, while others are at population level and so on. The relations between definitions of examination standards and the educational assessment paradigms introduced in Chapter 1 have also been outlined.

An essential dilemma for standard setting, which runs through the literature on definitions and methods, is the extent to which standards are evidenced by qualitative information about students' work or quantitative data. Put another way, are changes in examination standards best explained by students' performances or the difficulty of the examination? Cohort-referencing definitions do not have much to say about students' performances. In reaction to this, there was a criterion-referencing movement. In turn, this

approach was found to lack evidence regarding examination difficulty. The systemic definitions (attainment-referencing and standards-referencing) were introduced to tackle the need for information on students' performances *and* examination difficulty in standard setting. As outlined above, attainment-referencing and standards-referencing are the same at one level. They differ in terms of the underpinning philosophies and standards setting methods associated with them. Attainment-referencing arose from the curriculum-based examining tradition while standards-referencing arose from the psychometrics tradition. For the first time to our knowledge, the research on examination standards definitions reaches beyond a single country or a comparison of a small number of countries. Here, we showed that criterion-, cohort-, construct-, attainment- and standards-referencing were the definitional approaches used in the examination systems participating in our project. Exactly how these are enacted varied enormously, as depicted in the case study chapters. Although the research literature is dominated by psychometrics approaches, national examination standards across these countries derive from a range of perspectives in terms of their definitions of standards. The next chapter tackles the issues of context in much more detail.

References

Andrich, D. (2004) 'Controversy and the Rasch model: A characteristic of incompatible paradigms?'. *Medical Care*, 42, 7–16. Reprinted in Smith, E.V. and Smith, R.M. *Introduction to Rasch Measurement: Theory, Models and Application*. Maple Grove, MN: JAM Press, 143–66.

Andrich, D. (2018) 'Advances in social measurement: A Rasch measurement theory'. In Guillemin, F., Leplège, A., Briançon, S., Spitz, E. and Coste, J. (eds) *Perceived Health and Adaptation in Chronic Disease*. London: Routledge, 66–91.

Baird, J. (2007) 'Alternative conceptions of comparability'. In Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority, 124–56. Online. https://goo.gl/8SvTBo (accessed 7 June 2018).

Baird, J. (2010) 'What constitutes legitimate causal linking?'. *Measurement: Interdisciplinary Research and Perspectives*, 8 (4), 151–3.

Baird, J., Cresswell, M. and Newton, P. (2000) 'Would the real gold standard please step forward?'. *Research Papers in Education*, 15 (2), 213–29.

Baird, J. and Gray, L. (2016) 'The meaning of curriculum-related examination standards in Scotland and England: A home–international comparison'. *Oxford Review of Education*, 42 (3), 266–84.

Barker, R.G. (1968) *Ecological Psychology: Concepts and methods for studying the environment of human behavior*. Stanford, CA: Stanford University Press.

Blömeke, S. and Gustafsson, J.-E. (eds) (2017) *Standard Setting in Education: The Nordic countries in an international perspective: Methodology of educational measurement and assessment*. Cham: Springer.

Bronfenbrenner, U. (1974) 'Developmental research, public policy, and the ecology of childhood'. *Child Development*, 45 (1), 1–5.

Bronfenbrenner, U. (1976) 'The experimental ecology of education'. *Educational Researcher*, 5 (9), 5–15.

Bronfenbrenner, U. (1994) 'Ecological models of human development'. In Husen, T. and Postlethwaite, T.N. (eds) *International Encyclopedia of Education*. Vol. 3. 2nd ed. New York: Elsevier Science, 1643–7.

Christie, T. and Forrest, G.M. (1981) *Defining Public Examination Standards* (Schools Council Research Studies). Basingstoke: Macmillan Education.

Cizek, G.J. (1993) 'Reconsidering standards and criteria'. *Journal of Educational Measurement*, 30 (2), 93–106.

Cizek, G.J. and Bunch, M.B. (2007) *Standard Setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.

Coe, R. (1999) 'Changes in examination grades over time: Is the same worth less?'. Paper presented at the British Educational Research Association Annual Conference, University of Sussex, 2–5 September. Online. https://goo.gl/RhCtmt (accessed 19 June 2018).

Coe, R. (2007) 'Common examinee methods'. In Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority, 331–67. Online. https://goo.gl/VvyVHG (accessed 19 June 2018).

Coe, R. (2010) 'Understanding comparability of examination standards'. *Research Papers in Education*, 25 (3), 271–84.

Cresswell, M.J. (1987) 'Describing examination performance: Grade criteria in public examinations'. *Educational Studies*, 13 (3), 247–65.

Cresswell, M.J. (1994) 'Aggregation and awarding methods for national curriculum assessments in England and Wales: A comparison of approaches proposed for Key Stages 3 and 4'. *Assessment in Education: Principles, policy & practice*, 1 (1), 45–61.

Cresswell, M.J. (1996) 'Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches'. In Goldstein, H. and Lewis, T. (eds) *Assessment: Problems, developments and statistical issues*. Chichester: John Wiley, 57–84.

Cresswell, M.J. (1997) 'Examining Judgments: Theory and practice of awarding public examination grades'. Doctoral thesis, University of London Institute of Education. Online. http://discovery.ucl.ac.uk/10019189/1/322056.pdf (accessed 19 June 2018).

Gibson, J.J. (1979) *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Goldstein, H. (1979) 'Consequences of using the Rasch model for educational assessment'. *British Educational Research Journal*, 5 (2), 211–20.

Goldstein, H. and Wood, R. (1989) 'Five decades of item response modelling'. *British Journal of Mathematical and Statistical Psychology*, 42 (2), 139–67.

Hambleton, R.K. and Pitoniak, M.J. (2006) 'Setting performance standards'. In Brennan, R.L. (ed.) *Educational Measurement*. 4th ed. Westport, CT: Praeger, 433–70.

Hodgson, A. and Spours, K. (2015) 'An ecological analysis of the dynamics of localities: A 14+ low opportunity progression equilibrium in action'. *Journal of Education and Work*, 28 (1), 24–43.

Holland, P.W. (2007) 'A framework and history for score linking'. In Dorans, N.J., Pommerich, M. and Holland, P.W. (eds) *Linking and Aligning Scores and Scales*. New York: Springer, 5–30.

Lawn, M. (ed.) (2008) *An Atlantic Crossing? The work of the International Examination Inquiry, its researchers, methods and influence* (Comparative Histories of Education). Oxford: Symposium Books.

McLeroy, K.R., Bibeau, D., Steckler, A. and Glanz, K. (1988) 'An ecological perspective on health promotion programs'. *Health Education Quarterly*, 15 (4), 351–77.

McNamara, T.F. (1997) '"Interaction" in second language performance assessment: Whose performance?'. *Applied Linguistics*, 18 (4), 446–66.

McNamara, T.F. (2007) 'Language testing: A question of context'. In Fox, J., Wesche, M., Bayliss, D., Cheng, L., Turner, C. and Doe, C. (eds) *Language Testing Reconsidered*. Ottawa: University of Ottawa Press, 131–7.

Newton, P. (1997a) 'Examining standards over time'. *Research Papers in Education*, 12 (3), 227–47.

Newton, P.E. (1997b) 'Measuring comparability of standards between subjects: Why our statistical techniques do not make the grade'. *British Educational Research Journal,* 23 (4), 433–49.

Newton, P.E. (2003) 'The defensibility of national curriculum assessment in England'. *Research Papers in Education*, 18 (2), 101–27.

Newton, P.E. (2005) 'Examination standards and the limits of linking'. *Assessment in Education: Principles, policy & practice*, 12 (2), 105–23.

Newton, P.E. (2010a) 'Contrasting conceptions of comparability'. *Research Papers in Education*, 25 (3), 285–92.

Newton, P. (2010b) 'Thinking about linking'. *Measurement: Interdisciplinary Research and Perspectives*, 8 (1), 38–56.

Newton, P.E. (2011) 'A level pass rates and the enduring myth of norm-referencing'. *Research Matters*, Special Issue 2, 20–6.

Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (eds) (2007) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority.

Panayides, P., Robinson, C. and Tymms, P. (2010) 'The assessment revolution that has passed England by: Rasch measurement'. *British Educational Research Journal*, 36 (4), 611–26.

Payne, M. (2014) *Modern Social Work Theory*. 4th ed. Basingstoke: Palgrave Macmillan.

Sadler, D.R. (1987) 'Specifying and promulgating achievement standards'. *Oxford Review of Education*, 13 (2), 191–209.

Stokols, D. (1996) 'Translating social ecological theory into guidelines for community health promotion'. *American Journal of Health Promotion*, 10 (4), 282–98.

Wiliam, D. (1996) 'Standards in examinations: A matter of trust?'. *Curriculum Journal*, 7 (3), 293–306.

Wolf, A. (1995) *Competence-Based Assessment*. Buckingham: Open University Press.

Wolf, A. and Cumming, J.J. (2000) 'The inside story: The reality of developing an assessment instrument'. *Studies in Educational Evaluation*, 26 (3), 211–29.

Zumbo, B.D. (2007) 'Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going'. *Language Assessment Quarterly*, 4 (2), 223–33.

Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Olvera Astivia, O.L. and Ark, T.K. (2015) 'A methodology for Zumbo's third generation DIF analyses and the ecology of item responding'. *Language Assessment Quarterly*, 12 (1), 136–51.

Chapter 15

Culture, context and controversy in setting national examination standards

Tina Isaacs and Kristine Gorgen

Introduction

In the last chapters we looked at what standard setting encompassed in different jurisdictions and settings in both their meaning and their practical manifestations. Three models, or paradigms, were promulgated – psychometric, outcomes-based and curriculum-based – as idealized types of standard setting systems in order to elucidate the major different approaches to setting and maintaining standards.

As part of the Standard Setting Project, we wanted to explore how standard setting processes fit and work in the wider political, social and cultural context. We asked contributors about controversies and changes that had taken place in their jurisdictions, since these can be very helpful for illuminating wider context. Our findings showed that radical changes to assessment systems are very difficult to put in place, and therefore rarely occur. This chapter explores this issue and presents a framework to explain why deep-seated change is so rare. While it concentrates on curriculum-based exit examinations, it sometimes delves into wider curriculum and assessment issues.

The chapter begins with a scene-setting analysis of how accepted standard setting practices become enshrined through culture and context, concentrating on theorists who have grappled with the relationships between education and culture. It puts forward some examples of the ways different countries use national assessments and examinations in practice, many of which are drawn from the Standard Setting Project. Accepted standard setting practices have been subject to challenge, and in the face of those challenges some have undergone, or are undergoing, operational change. The catalysts for those changes are investigated, employing a theoretical

framework that acknowledges the seminal work of Thomas Kuhn on paradigm change, while concluding that changes in standard setting are more often accommodations to existing models (or paradigms) rather than what Kuhn labelled paradigm shifts.

Culture and context

Standard setting systems are embedded in a country's assessment system. They are affected by the wider assessment ethos and by cultural and contextual conditions within the country. Educational cultures differ across jurisdictions, permeating their assessment structures and processes in idiosyncratic ways. Different cultures and contexts give rise to a variety of accepted practices in national assessment and examination systems across the world and often act as impediments to change. Eminent Swedish educator Torsten Husen argued that 'any educational system can only be fully understood in the context of the culture, traditions, history and general social structure of the nation it is designed to serve' (Husen, 1967: 220). Much has been written about the global convergence of education policy (and assessment systems) and academics and interested parties, such as the OECD, have written extensively on the topic (see, for example, Meyer and Benavot, 2013; Morgan and Shahjahan, 2014). However, the discussion of standard setting systems remains a bastion of the local in our globalized assessment world. This chapter therefore offers a brief discussion of the ways in which culture and context shape assessment and standard setting systems before presenting a framework for standard setting system change.

The feedback loop of culture, context and education

This section starts with a brief general exploration of education and culture, then moves on to investigate the particular role of assessment in shaping and being shaped by context and culture, using examples from the Standard Setting Project.

Educational thinkers, including luminaries such as John Dewey and Jerome Bruner, have written about the relationship between education and culture. Dewey (1938) was concerned with educational culture, referring to established practices and patterns of organization and thought as inhibiting change and reform in schools. He argued that through the established educational culture a teacher 'could content himself with thinking of the next examination period or the promotion to the next class' (Dewey, 1938: 39), instead of investing more innovatively in the mental growth of students. Looking at culture and education not on a school level but in a wider societal context, Bruner (1996) stipulated that education was at

the same time born out of a society's culture and the instrument by which this culture is carried forth across generations. Bruner (1960) furthermore argued that well-designed examinations can be a helpful tool to improve curriculum and teaching, as well as assessing a student's progress.

One of the most prominent scholars on the relationship between culture and education (and assessment) was the sociologist and philosopher Pierre Bourdieu. Bourdieu (1990) saw education not just as an instrument to protect and maintain culture, but also to ensure the reproduction of social inequalities. Bourdieu argued that, since elites design and control the education system, it is the children of these elites who will be successful in it, which in turn legitimizes their elite status in the future. Bourdieu paid particular attention to the role of examinations in social and cultural reproduction. Bourdieu (1990) saw culture as one of the 'factors that can explain the historical or national variations in the functional weight of the examination within the education system' (Bourdieu, 1990: 144). If education is at the same time born out of culture and an instrument to safeguard it for the future, national examinations are a means to standardize and monitor the interplay of education and culture.

Baird and Gray (2016) found that cultural context and its effect on attitudes to examination results was key in determining what would be accepted and what was controversial in a standard setting system. Even between Scotland and England, two parts of the devolved education system of the United Kingdom, they found striking differences. The cultural position of the examination system in Scotland was found to have inclusion at its heart, while that in England was found to be more elitist, emphasizing the selective function of examinations. These differences in cultural context affected what was considered important in both England and Scotland and thus influenced standard setting concepts and methods.

Many of the Standard Setting Project's jurisdictions use curriculum-related examinations to select learners for higher education, work and other study options. Such examinations are employed for school leaving certification, or university entrance, or both. They are also used for school accountability, to measure system performance, or to allocate resources, reflecting the perceived needs of governments, higher education, employers and society at large.

These uses are not uncontested, as explored below. Culture and context can influence the way people in different jurisdictions understand and place value on examinations. They might ask whether or not examinations are testing what ought to be tested and may come up with myriad answers to questions such as: can tests be fair and equitable to all test-takers regardless

of ethnicity or social and economic status? Do tests adequately measure what students should know, be able to do and be like in the twenty-first century? Are examinations capable of allowing judgements to be made about students' performance over time or about their communication and problem-solving skills? Do timed tests allow students to demonstrate the processes by which they have developed their thinking, arrived at their answers and planned their work? Can, and should, examinations assess an entire curriculum? How these questions should and would be answered depends on what a society wants from its assessment system and what the balance should be between judgements based on examination outcomes and teacher judgement.

Non-examined, school-based assessment, on the other hand, ostensibly allows teachers to assess the implemented curriculum and provides a more nuanced look into students' skills, knowledge and understanding, using a variety of assessment instruments such as rich tasks, projects and portfolios. Queensland and Sweden (see Chapters 10 and 12), for example, both place a high value on teacher judgement and on continuous assessment. However, assessment experts have warned that reliable and valid school-based assessments are difficult to design and can fall foul of both construct irrelevance and under-representation – that is, assessing things that are not part of the curriculum or neglecting to cover the whole curriculum (Black *et al*., 2011; Gardner *et al*., 2010; Harlen, 2005; Klenowski, 2009; McCann and Stanley, 2010; Stanley *et al*., 2009). It is also difficult to reliably ascertain differences in performance across individuals who do the marking and grading – that is, inter-rater reliability. This is especially true in the US, where most states rely on teacher judgement in the overall grading of students. In-school and between-school moderation has been proposed as the best way of increasing reliability.

Having briefly looked at culture and context in general, we turn our attention to how and why controversies in examination and standard setting can lead to assessment system change, either fundamentally or, more often, around the edges. Although much of the Standard Setting Project focused on the more technical aspects of setting and maintaining standards, especially the assessment and standard setting processes, in order to help understand the culture and context of particular standard setting policies, each in-country expert in our 12 case study jurisdictions was also asked to elucidate some of the political and public controversies and debates concerning their system's school leaving and university entrance examinations. They were asked to analyse, in the context of the most recent assessment reforms: what the main political and public controversies and debates about examinations

standards were; what research evidence existed relating to these debates and controversies; and what the controversies tell us about the political and public views of examination standards. Through the elucidation of political and policy debate, we were struck by how rare deep-seated, fundamental reform to standard setting practice is, and started to think about why. The next sections explore this and offer a framework to explain change (or lack of it) in assessment practice.

Controversy and contention in examination standards

Structuring changes to standards

In Chapter 1, we presented the notion of idealized type paradigms in educational assessment – psychometric, outcomes-based and curriculum-based. In exploring why certain jurisdictions attempt to change their assessment system, sometimes contemplating a shift from one educational paradigm either partially or wholly to another – and why so few succeed in doing so in a fundamental and deep manner – both Thomas Kuhn's and Michael Fullan's ideas are helpful in explaining the issues.

As explained in Chapter 1, Kuhn (1962) defined a paradigm as something that offered a 'universally recognised scientific achievement that, for a time, provides model problems and solutions for a community of researchers' (Kuhn, 1962: 10). Paradigms are guides to what phenomena or attributes should be observed and studied, what kinds of questions researchers might ask, how those questions should be structured and how investigation results should be interpreted. Mirroring what Kuhn described as a pre-paradigmatic state, in educational assessment there are coexisting, not always compatible and sometimes competing theoretical models in the applied fields of psychometric, outcomes-based and curriculum-based assessment across jurisdictions and countries. Chapter 1 also observed that in the social sciences sphere, social and political forces play a much stronger role than in Kuhn's physical sciences arena. Worldviews are much more contested in the social sciences due to their social and political contexts. While recognizing that in practice there are almost no pure examples of the three models in high stakes, end of school examinations, distinctions were drawn between them to show their distinctive traditions and assumptions.

One of Kuhn's most lasting contributions was his analysis of how accepted scientific paradigms might be superseded by other paradigms – what he termed paradigm shifts. For Kuhn (1962), paradigm shifts start when anomalies arise that cannot be easily resolved within the existing, accepted, scientific paradigm. These anomalies lead to a re-evaluation of existing data and theories. When there are enough anomalies, and there is

an alternative explanation that proves more compelling to enough people, paradigm shifts occur.

As outlined above, just like scientific communities, cultural communities might have accepted practices and distinctive traditions in their education and assessment systems. If there is reason for enough people to doubt the proper functioning of a system, this might lead to a re-evaluation of the accepted practices. At the same time, the culture and context that frame accepted practices in examination systems might change, leading to a tension between the established practice and the newly held beliefs, values and social needs. Kuhn's ideas about scientific paradigm shift are helpful in investigating the very different world of assessment paradigms, as they help to explain why major reforms in assessment systems are rare.

Although the applicability of Kuhn's work to the social sciences has been questioned and Kuhn himself was cautious about extending his ideas beyond the history of natural science (1962), we found his ideas a useful launching point when trying to understand the impetus for change in the context of standard setting. Using Kuhn's ideas more generally, we developed a framework for understanding impulses behind the desire for paradigm change, starting with three required preconditions:

- condition 1: there must be dissatisfaction with the current, accepted paradigm
- condition 2: there must be an alternative, agreed upon, paradigm that is a better fit
- condition 3: advocates of the new paradigm must outnumber or outweigh those supporting the old paradigm.

Only once all three of these conditions are met can paradigm shift occur, but as we will see, accommodation within the accepted paradigm is more likely than the adoption of a different one.

What might these conditions look like in standard setting? Taking condition 1 first, it is clear that 'anomalies' in the sense suggested by Kuhn are not wholly applicable in the context of educational assessment. The three idealized types as outlined in Chapter 1 can exist simultaneously, and many assessment systems are comprised of elements of one or more of the models – even if one of the models predominates. This, of course, does not mean that standard setting systems remain static, and we suggest that dissatisfaction with the accepted assessment model is exacerbated by controversy that is either substantial or prolonged. We have identified four elements that can lead to such controversy and analyse them more fully later in the chapter:

- *examining crises*, that is, dissatisfaction can arise as a result of problems with the examining process itself: question paper errors, widespread issues with results, failure to despatch results on time and so on. Examining crises can either be the result of genuine failings and errors or of the perception that a failing or error has been made
- *media reporting*, that is, when issues having to do with standard setting break loose from the closed domain of ministries, regulators and examination boards and are brought to light by the social, print and broadcast media, often blowing them out of proportion or simplifying them in ways that make the problems seem far greater than they are
- *political involvement*, often in tandem with media reporting: this is when politicians use standard setting and assessment issues as ways of promulgating political agendas, mainly around the meaning of standards
- *sociocultural drift*. Societal and cultural views change over time calling into question the accepted standard setting practices of a jurisdiction. Dissatisfaction may arise when sociocultural values and goals change to the point that the current system is no longer aligned with the values and goals of individuals within the system.

In practice, it seems likely that all four elements, which are often inextricably linked, would be required to generate sufficient discontent for fundamental change to occur. Sociocultural drift – or more specifically, misalignment with cultural values – may determine which examining crises gain enough attention to inspire widespread dissatisfaction. For example, as Baird and Gray (2016) noted, significantly increasing pass rates was far more controversial in England than in Scotland as a result of different sociocultural values and expectations. While the media and politicians can, to some extent, manufacture controversies (see Baird *et al*., 2011, 2016; Murphy, 2013), unless the controversy 'rings true' to system users, it is unlikely to generate enough unhappiness for a deeply rooted rethink. Similarly, some basis in fact – such as a genuine (or widely perceived as genuine) examining crisis – is needed for the controversy to be compelling.

However, while changes to scientific paradigms need only convince scientists in the relevant field in order to be effected, educational systems exist in a more complex ecosystem (Sirotnik, 1998, 2005), and must convince a far wider range of stakeholders. The involvement of the media is essential in doing this. Finally, in most countries, significant changes to education or assessment cannot take place without the sanction of politicians, who may also play a role in generating or inflaming dissatisfaction.

Condition 2, the existence of an agreed alternative paradigm, is necessary for paradigm shift as it directs widespread dissatisfaction towards a seemingly better option. In scientific scenarios, this means it must provide a better explanation for the existing body of knowledge. In standard setting, a paradigm can be viewed as 'a better fit' if it addresses the dissatisfactions with the current system. We found this condition to be the most challenging of our framework (see below).

Finally, condition 3 requires that advocates or practitioners of the new paradigm prevail over those of the old paradigm. Achieving this condition in a standard setting context is closely linked to the mechanisms for reaching sufficient dissatisfaction outlined for condition 1. However, unless those who hold the genuine power in the education and assessment system convert to the new paradigm and adjust their behaviour accordingly, any changes will be at surface-level only: a phenomenon reported by Ball *et al.* (2012) in their work on policy enactment. Accommodation within the current paradigm or a retreat to former standard setting methods often results.

This interpretation of Kuhn's work fits neatly with the arguments of Michael Fullan (2005, 2006, 2016), whose work is more focused on the successful implementation of educational change than on assessment systems, per se. Fullan (2016) stresses that effective change means shaping and reshaping good ideas, building capacity and ownership in stakeholders. He suggests (2014) that individuals are the core unit of change: if they lack alignment with the goals and values of the proposed change (condition 3), or if they lack the skills to implement the change (condition 2), then change will not be implemented successfully, or sometimes at all. The wrong drivers – external accountability and fragmented strategies – also undermine change. Change is successful when it comes about through the aggregate efforts of large numbers of individuals working towards the same goal (condition 1), in a way that engages all those whose 'buy-in' is required (condition 3). A critical mass of people who are skilled in and committed to the change must be generated and the system has to continually support all those working within it. 'Higher, clearer standards, combined with correlated assessments, are essential along the way, but they will not drive the system forward' (Fullan, 2016: 43). And even when changes are introduced and implemented, they are often discarded or abandoned – 'we might assume that specific educational changes are introduced because they are desirable according to certain educational values and meet a given need better than existing practices do ... however, this is not the way it always or even usually happens' (ibid.: 59). He notes that innovators sometimes do

not take enough account of the larger cultural picture or how people will react to their proffered reforms, and argues that successful change requires recognizing and working through conflict and disagreement, and having enough time to get rid of barriers. In sum, there is clear overlap with our interpretation of Kuhn's preconditions for paradigm change: there must be sufficient dissatisfaction with the current paradigm that a large number of individuals feel inspired to change (conditions 1 and 3), and there must be a workable alternative (condition 2).

We must be mindful, though, that in education systems, radical shifts – or attempts at them – do not always work out as intended. 'Change often has unexpected consequences, not least in education policy. The nature of policy implementation means that intentions do not always translate into the expected outcomes. Policymakers cannot always pre-guess the cultural influences of policy once they have passed through the boundaried institutions of school, college or universities' (McCaig, 2003: 487). And, as Baird and Opposs stress in Chapter 1, in practice, even borrowing methods and ideas across standard setting paradigms can cause significant strains because of the very different beliefs that underpin each model.

Employing the change framework

We posited above that there are four major catalysts that might trigger attempts to shift standard setting paradigms, either partially or fully: examining crises; media attention; political involvement; and sociocultural drift. We explore each in turn below – although they do, of course, overlap – largely using evidence from the Standard Setting Project. They provide empirical evidence of the circumstances that trigger condition 1 of our framework and to a lesser extent condition 3. These should provide impetus for fundamental systemic change, except that condition 2 is rarely present – there is little consensus among stakeholders (politicians, education professionals, students, parents, the general public) that a different assessment paradigm will solve their problems or even what that paradigm should be. To elucidate further, we now turn to our four catalysts.

Examining crises

Kathleen Rhoades and George Madaus (2003) point to what they call the 'wilful ignorance' (9) to which many involved in high stakes testing adhere. They blame this phenomenon for a misguided belief that testing is a precise science. Rhoades and Madaus catalogue a litany of human errors in examining – as opposed to measurement error, which is unavoidable – mostly across the US, but also elsewhere in the world that have undermined earlier faith in the examining system, providing good examples of condition 1

of our framework. They contrast 'active' human error – one-off mistakes by individuals – with 'latent' human error, which is caused by 'misguided executive decisions' by examination boards and policymakers that have the capacity to cause multiple and serious active errors (ibid.: 6). It is these latent errors that sometimes act as change catalysts but that much of the time create short-lived public ire and then are swept under the carpet. Latent errors include educational assessment legislation that flies in the face of expert advice, policymakers' insistence on examination boards working to impossible timescales, the belief that examination outcomes should improve (or conversely stay the same or decline) over time, and the misuse of examination results as the sole arbiters of student achievement.

One of the major examples of latent error that Rhoades and Madaus (2003) highlight is that of the Scottish examinations in 2000, in which over 5,000 potentially university bound students ostensibly received incomplete or inaccurate results. Media and political debate ensued. Appeals skyrocketed. Rhoades and Madaus attribute the problems to lack of resources – both fiscal and human – poor planning and a very tight timeline. Baird and Lee-Kelley (2009: 58), after studying a major review of the events, attribute the difficulties to a lack of planning and monitoring, leadership and delegation problems, low level of management skills and politically driven changes without scoping of projects.

England's Curriculum 2000 A level 'crisis' mirrored the Scottish one. New qualifications were put in place for teaching in 2000, and there was widespread concern about the reliability of the grades in 2002. This provoked a review of A level grading that pointed out that changes to standard setting and grading procedures had been rushed, comparability with the older version of A levels had been compromised, and communications with teachers and students had been poor (Tomlinson, 2002). There was policy pressure not to have the 2002 results too dramatically out of line with the 2001 results. In the ensuing action, more than 90,000 examinations were re-marked, and although most did not result in overall A level qualification grade changes, over 100 students initially missed out on their university places (Tomlinson, 2002). Taylor and Opposs explore more recent controversies in English examinations in Chapter 6.

David Lines (2000) explained the problem in a way related to Rhoades and Madaus's (2003) wilful ignorance, stating that the English examinations system was:

> built on the erroneous assumption that external examinations are accurate, fair and efficient, while assessment by teachers is

> not. This notion has been brought about by the determination of successive governments to centralize and control all aspects of education ... we have an examinations industry ... shorn of old standards and values, but required to serve increasing numbers of demanding customers. It is hardly surprising that accidents happen (Lines, 2000: 1).

However, the Curriculum 2000 crisis in England resulted in changes that were more aligned to accommodation than paradigm shift. Lines's call for a system based on teacher assessment (a shift from the curriculum to the outcomes-based paradigm) was for the most part ignored.

In Ireland, where grade boundaries are fixed and raw scores not standardized, leading to the standard setting process being contained within marking, concern has been expressed that the rise in the proportion of students receiving high grades has not been matched by learning improvements. Chief examiners are also concerned that using mark schemes to regulate standards and keep the grade distribution stable over time undermines potential innovation and the examination of higher order thinking skills as well as promoting 'gaming the system' (see Chapter 9). It is too early to know whether these concerns will lead to substantive changes to the standard setting system or changes from one assessment paradigm to another; dissatisfaction could lead to a shift in the system or simply to accommodation.

McCaig (2003) reminds us that assessment crises 'tend to be time-limited as the results-reporting stage fades in the memory and there is plenty of time for any necessary reforms to be carried out before "next summer"' (McCaig, 2003: 472) (or at the end of the next examination cycle) and that 'exam crises by their very nature are seasonal in that they can be expected to have a limited life in terms of public opinion and media interest' (ibid.: 473). However, when crises are substantial or deep, as in the Scottish and English crises of 2000, their memory fades considerably more slowly, while at the same time, those responsible have a hard time – either through lack of resources or lack of political will – to carry out 'necessary reforms'. In such a case condition 1 may be fulfilled, but conditions 2 and 3 are negligible or even absent.

Media attention

Tumultuous media attention compounds examining crises, and can act as a change catalyst, with front-page stories seemingly going on for days, especially during slow news seasons such as the summer. This press barrage serves to help undermine the trust that teachers, students and parents

might have in the examination system, despite any public reassurance from governments, regulators and examination boards that the problems may in fact have been overplayed and, in any case, will be shortly remedied.

Since standard setting and maintenance are very complex and technical, and not well understood publicly, any media insinuation that something is amiss can cause deterioration in confidence, especially in an era when the media are increasingly paying more attention to the technicalities (Billington, 2006). Negative media coverage can cause the public to doubt the validity and fairness of examination standards. Newton (2005), believing that the media are the largest impediment to public understanding of examinations, advocates for more information and transparency from examinations developers.

Sometimes transparency can have a different effect from what was intended. Since the advent of the national curriculum in Australia, more and more information has been put in the public domain. In Queensland the media uses official – and unofficial – data on student attainment to produce performance tables that compare achievement between schools, which according to Campbell (Chapter 10) has redirected schools' efforts toward improving that which is reportable. They believe that this may have resulted in more student preparation for Queensland's Core Skills Test at the expense of more subject-based teaching and assessment.

Political involvement

With the advent of international testing such as Progress in International Reading Literacy (PIRLS), Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA) and their concomitant rank-order tables and heightened media attention, and seemingly sharing a belief that an educated populace brings economic prosperity and a larger slice of a finite global pie, governments seek to measure and improve their education systems' quality and results (there is a huge literature on the impact of international testing, to which this chapter cannot do justice). Education policy objectives include both direct and indirect intervention into standard setting and maintaining. Many policymakers believe that examinations can be progressive, equitable, rational and reasonable and are a valuable tool for education reform (Scott, 2011). Politicians, often needing to show positive educational outcomes between elections, cannot give adequate time for reforms to bed in. They might believe that the examinations are the easiest part of the system to change and use examination reform as a level to effect other changes that they find more intractable. Governments develop feedback and accountability

systems accordingly, which results in examinations becoming both a policy instrument and a tool to ascertain policy effectiveness (Herman and Baker, 2009). There is also the issue of wilful government and media misrepresentation of examinations outcomes to support existing policies or to make a case for policy alteration (Billington, 2006; Mansell, 2013).

Government influence on the characteristics of examinations is not uncommon. In Sweden, when government is led by conservatives, there is more of an appetite for national tests, school monitoring and earlier grading of students (Chapter 12). Balancing a system that relies on teacher judgement with the desire to maintain comparable standards is proving tricky. In Chapter 12 Wikström and Pantzare point to political and public considerations on how to maintain the teacher assessment model while alleviating the reliability and comparability issues associated with it. They argue that the current grading standards system cannot both provide apposite information on what students know and are able to do upon leaving secondary education *and* impart technically robust information on school performance over time and act as a university selection tool. While there have been calls for the increased use of national tests, their reliability, too, has been questioned. Gustafsson *et al*. (2014) have called for the creation of new tests.

In Sweden the debate is not so much about how standards are set as about who will be setting them. At the moment the system is based primarily on teacher input (outcomes-based paradigm) – it is teachers who have responsibility for assessing students based on internal (classroom-based) evidence. Social resistance to high stakes standardized testing is formidable, seeming to overcome reservations about unreliable teacher grading, although unreliability issues, as well as concerns about grade inflation, have resurfaced lately (Chapter 12). Erikson (2017) writes about a recent National Agency for Education inquiry into the national assessment system, the results of which were published in the spring of 2016. It concluded that national tests should be mandatory, complemented by national support materials as well as a national evaluation system. It also proposed a clearer and more robust relationship between national test outcomes and grades.

Changes to the party in power can sometimes forestall what seems to be a fundamental change to the standards system. In the US two consortia were commissioned to develop assessments to test the Common Core State Standards, which had been generally accepted by the states when they were introduced in 2010. However, both the Common Core and its assessments have much fallen out of favour in the current political climate in the US, and more and more states are abandoning this shared curriculum. As Morgan

asserts (in Chapter 13) the Advanced Placement programme contains almost the only US national curricula and assessments.

Sociocultural drift

As outlined above, changes in culture and context can serve as a catalyst for change. What follows concentrates on sociocultural factors that directly affect standard setting in exit examinations and therefore university entrance. If, for example, an increasing number of students want to enter university after secondary education, this might require the examination system to change in a way that ensures that university entrance decisions can be made. It could also be the case that in pursuit of social justice a country changes its legislation for compulsory education, encouraging girls to complete secondary education or racially integrating schools. This sort of shift in the country's education system might require assessment system reform. Earlier in the chapter we alluded to perceptions of standards being lowered and examinations dumbed down; the debate that follows is whether standards *should* change in the face of fairness and social justice, drawing on examples from Chile, Georgia and South Africa.

Both Chile (Chapter 5) and South Africa (Chapter 11) underwent massive cultural changes after emerging from dictatorial and apartheid regimes; their examination systems attempt to reflect issues of social justice and the end of socio-economic and racial or ethnic segregation. One of the arguments made in favour of the Chilean university entrance tests (the PSU) was that because they were curriculum-based they would promote equity – a promise the examinations have not seemed able to keep (see Chapter 5). Georgia (Chapter 8) brought in centrally administered unified standardized university entrance examinations to provide students with access to higher education that is free from the corruption endemic in the past. Georgia gives very little money to the university sector, which means that 90 per cent of university revenues come from student fees. This has interestingly brought about what Andguladze and Mindadze in Chapter 8 state is an assessment system without standards – if the government had introduced stringent standards for university entrance at least 30 per cent of current students would not have got places, which could potentially mean the closing down of some less competitive colleges and programmes.

Georgia's open-ended university admissions system has meant that the standards of the Unified National Examinations take on a different connotation from those systems where university selection is highly competitive. Andguladze and Mindadze argue that the key to understanding the relationship between examination standards and university entrance in

Georgia is that one has to take into account the country's current policy goals. Since 2005, admission to higher education has been organized centrally based on three compulsory examinations and one elective examination. The result of these examinations is the only criterion for admission to all private and public higher education institutions in Georgia and for student funding decisions. Chankseliani (2013) argue that the new admission system might have contributed to limiting corruption. However, since no other information about the applicant – such as family background, gender, or minority status – is taken into account, the centralized admission system disadvantages rural and less privileged students and has thus replaced corruption with systemic reproduction of inequalities. Georgia's university-bound population suffers from an overall lack of readiness for the complexity of university programmes, and universities must develop strategies for integrating these students. Andguladze and Mindadze point to OECD findings that over half of Georgian 15-year-olds get PISA scores below level 2 – a low level for functional literacy in reading, mathematics and science. Given these deeply rooted challenges, they argue that it is surprising that more attention is not paid to standard setting and the role of curriculum-based assessments to help pinpoint where reforms could be made.

Increasing the number of disadvantaged students who can access higher education is a goal of the Chilean government; In Chapter 5 Osses and Varas argue that the current reliance on the outcomes of the PSU might be impeding that. Academically selective universities make use of the PSU in setting minimum admissions requirements, but less selective tertiary programmes have opened to freer access arrangements over the last few years through government-initiated programmes. University funding for disadvantaged students is supposed to be tied to the use of fair and transparent selection procedures, and Osses and Varas contend that this warrants an expansion of the current university entrance testing regime. In this context, Chile had hoped that the introduction of a new form of curriculum-based university entrance examination, the PSU, would increase fairness to those from lower socio-economic groups. It is questionable, however, whether or not the gap narrowed. While the university council of rectors (CRUCH) claims that the PSU modestly decreased the gaps between socio-economic groups, Osses and Varas argue that because the examinations only assess the general curriculum, those on vocational tracks, who are generally socio-economically deprived, are disadvantaged. They advocate a new set of instruments that can assess students from a wide range of backgrounds and that also incorporate the needs of those who are on vocationally oriented upper secondary programmes.

It is almost 25 years since the end of apartheid in South Africa, and successive governments have struggled with how successfully to balance secondary school standards with widening participation. Sayed and Ahmed (2011) point to the many challenges South Africa faces in its attempt to combine equity and quality in education. These challenges, the result of decades of oppression, are local as well as global; they are to be found in other developing countries in the context of globalization. Kanjee and Sayed (2013) point out that the assessment practices in South Africa since the end of apartheid in 1994 have focused on empowering the previously disadvantaged black population through the introduction of an outcomes-based model. However, they also refer to policy imperatives that have favoured the retention of traditional, measurement-based forms of teaching and assessment (see Chapter 11 and below).

Meeting the conditions for paradigm shift

The examples above illustrate an abundance of challenges to accepted practice and often frustration about how to deal with it. Therefore condition 1, dissatisfaction, of our framework is fairly easily met. Not so condition 2, an alternative, or condition 3, support for change. While different idealized-type models for standard setting exist, as evidenced by the psychometric, outcomes-based and curriculum-based archetypes, and could be characterized as being fundamentally different from each other, stakeholder consensus that there is a better fitting model to the 'accepted' one is often lacking. Advocates for change must not only outweigh those who support the current practices, but must also prevail in an often politically charged atmosphere, which by its very nature can be slow to change. Standard setting is an activity that takes place within national education policymaking contexts, which naturally sets it somewhat apart from Kuhn's scientific activity. Transforming standard setting processes is not directly akin to a scientific community coming to recognize that its own assumptions are breaking down. Instead it is about pressures and controversies developing somewhere in a disparate set of political, technical and social interests, which often leads to expedient decisions being taken by policymakers or the education public servants who work for them. Scientific 'revolutions' suggest a rationality and use of evidence that may be less manifest in debates about examinations. Agency is a key consideration here: often the power resides in people and institutions that are less 'expert' and more political than the scientific communities that Kuhn describes.

In addition, as we have seen, psychometric, outcomes-based and curriculum-based archetypes exist simultaneously, sometimes within the

same standard setting system, as in Sweden and Queensland. In the face of large-scale dissatisfaction and even crisis it is rare to find a system that shifts from one standard setting model to another. Probably New Zealand went the furthest in shifting from one type to another – from a system based on norm-referenced examinations to one based on outcomes, starting in the 1980s.

New Zealand shifted to a modular, outcomes-based system in its senior secondary schools due to mounting dissatisfaction among teacher groups, employers and politicians, who were concerned that the examination system was outdated and could not fulfil society's needs during a time of rapid economic change (Lee *et al.*, 2013). This development was part of the introduction of a national qualifications framework in 1991 that included both vocational and academic qualifications. The University Entrance Examination was replaced first by the internally assessed School Certificate and then the National Certificate of Educational Achievement (NCEA) between 2002 and 2004. New Zealand adopted this most far-reaching outcomes-based approach in order to become more globally competitive, accountable and rigorous, while at the same time bridging the academic-vocational divide. Lee *et al.* (2013) ascribe the relative ease by which the system's change took place to government officials 'committed to radical changes in senior secondary school curriculum and assessment' (Lee *et al.*, 2013: 38).

Secondary school qualifications were developed that contained unit and achievement standards derived from New Zealand's national curriculum and include learning outcomes and assessment criteria (standards). They are available at three levels, each of which is primarily aimed at students in Year 11 (level 1), Year 12 (level 2) or Year 13 (level 3). Assessment is both teacher- and externally based. The system has been criticized for intensifying assessment, fragmenting teaching and learning, increasing teacher workload and 'a potential dumbing-down of the curriculum associated with the aim of keeping more students at school' (Philips, 2006: 4). The NCEA was not piloted and consultation on it was ostensibly limited. Philips highlights the far-reaching nature of the changes:

> Generally speaking ... other countries have not tended to adopt the same radical reform as in New Zealand, preferring instead to take a more cautious 'incrementalist' approach, such as in the various countries making up the United Kingdom, and the various states in Australia (Philips, 2006: 6).

Policy objectives included recognizing a wider range of achievement, addressing major demographic changes and providing a foundation for economic and social development (Philips, 2006).

Concern has been expressed about whether the performance-based learning system is sufficiently robust for university entrance. Competitive university programmes sometimes either imposed additional requirements alongside those needed for general entry or they specified courses that applicants should take while in upper secondary. Some universities delayed selection to competitive programmes until a student's second year of university so that they could use results from the first year in the decision-making process (Vallender, 2009). Some secondary schools switched to Cambridge International Examinations, which were perceived to be a less easy option (Johnston, 2015; Vallender, 2009).

In 2004 there was a standard setting controversy about the processes to set and moderate standards for the externally set Scholarship Assessment, which gives a monetary award to the most able university applicants (Martin, 2005). The award was thought to have been insufficiently exemplified and comparability between subjects was questionable. A report into the award pointed to 'drift into implementation without adequate analysis of the strategic policy risks' (SSC, 2005: n.p.). Critics claimed the awards were unfair because the results were not scaled to ensure comparability across subjects and across years, which may have had negative consequences for some university applicants (Martin, 2005). This was compounded by the separation of standard-setting roles and lack of coordination between the Ministry and the New Zealand Qualification Authority (NZQA).

The controversy over whether the outcomes-based approach is fit for purpose continues in New Zealand. Recently there have been concerns about grade inflation – the outcomes of the NCEA have risen dramatically since 2004, while New Zealand's students' performance on PISA tests has declined (Powlesland, 2017). Powlesland (2017) claims that this decline in standards has meant that universities have had to raise their entry requirements. Boereboom (2016) points to the 'major paradigm shift in assessment for New Zealand school qualification from a norm-referenced system to a standards-based system' as the reason that the NZQA has had to reformulate university entrance requirements. Boereboom argues that with limited grades available and low floor standards in literacy and numeracy, universities have great difficulty in discriminating among applicants. He also bemoans the fact that internal and external assessments get equal weighting. While New Zealand seems to have no intention of returning to its former standard-setting system, the shift from the norm-referenced model to a

largely performance-based one continues to be contested, especially around issues of university entrance.

Following the lead of New Zealand, Scotland, some Australian states and Canadian provinces, South Africa introduced outcomes-based education (OBE) in 1995 as part of its immediate post-apartheid reforms (Jansen, 1998, 2002; Botha, 2002; Cross *et al*., 2002; Sayed and Ahmed, 2011; Kanjee and Sayed, 2013; Schmidt, 2017). One aim of the curriculum and assessment reforms was to:

> introduce a shift from a system that is dominated by public examinations, which are 'high stakes' and whose main function has always been to rank, grade, select and certificate learners, to a new system that informs and improves the curriculum and assessment practices of educators and the leadership, governance and organisation of learning sites (quoted in Kanjee and Sayed, 2013: 464).

Learner-focused and teacher-led assessment in outcomes-based education largely centres on formative (what became known in South Africa as 'informal') assessment in order for students to achieve set goals for particular learning phases. Learning areas replaced subjects, with concomitant specified outcomes, range statements and assessment criteria (Cross *et al*., 2002) (outcomes-based). Knowledge and skills were integrated, and the emphasis was on competency, knowledge and attitudes gained through teamwork, critical thinking and problem solving (Cross *et al*., 2002; Sayed and Ahmed, 2011). Formative assessment tasks were to be used to support students in a developmental manner, feeding back into teaching and learning. The policy commitment, however, did not translate into in-depth teacher training, and the implementation foundered (Schmidt, 2017). In 2000 a government review committee noted that OBE lacked 'alignment' between curriculum and assessment policy (Kanjee and Sayed, 2013: 450) and recommended that the curriculum be simplified and some specific curricular outcomes (and their associated standard setting mechanisms of assessment criteria, performance indicators and performance levels) dropped, although the notion of assessing students' performance against assessment standards of overall learning outcomes remained. Kanjee and Sayed (2013) see this as a shift from the 1998 criterion-referenced assessment to standards-referenced assessment. While the role of teacher-based assessment remained important in the 2007 revisions to the South African curriculum, references to OBE were dropped and guidance on 'formal', or summative assessments that are used for examinations purposes and that impact on pass/fail decisions,

enhanced. Importantly for the issues concerned in this book, the Grade 12 matriculation examinations, analysed in Chapter 11, retained their time-honoured importance, and in a return to the testing-oriented curriculum-based paradigm, formal assessments have been introduced for primary and lower secondary students. Kanjee and Sayed (2013) and Schmidt (2017) conjure up myriad reasons why the reforms failed to take hold: the poor quality of schooling; a lack of resources; large class sizes; too-strenuous demands on teachers; a paucity of effective training and guidelines; lack of assessment knowledge; the introduction of additional standardized testing and other progress measures; a lack of alignment between curriculum and assessment policy; willingness to listen to foreign consultants at the expense of local practitioners; over-hasty and uncritical policy borrowing; a lack of policy vision.

As Schmidt (2017) points out, many of the systems that introduced outcomes-based models have either abandoned the model or continue to suffer failure and intense criticism. Most systems accommodate and adjust; instead of structural revolution, we are more likely to find containment and adaptation. Although people may be dissatisfied with the current model of standard setting, as illustrated in the examples above, alternate models of standard setting lack the critical mass of support, and buy-in for systemic change is difficult to come by, making them unworkable in the jurisdictions' current context.

Summary

This chapter explored the ramifications of culture, context and controversy in the standard setting realm, concentrating on exit examinations, but sometimes delving into wider curriculum and assessment issues. Starting from Baird and Opposs's examination of the three idealized types (paradigms) of assessment we investigated why, in the face of dissatisfaction with the dominant paradigm, jurisdictions on the whole did not indulge in root and branch change from one idealized type of standard setting to another, but instead largely made accommodations within the dominant paradigm. In order to answer our questions about change we looked at theoretical underpinnings primarily as elucidated by Thomas Kuhn and Michael Fullan and produced a framework for change based on three conditions: (1) dissatisfaction with the prevailing paradigm; (2) the existence of an alternative paradigm that is a 'better fit'; and (3) the forces for change outweighing the desire for accommodation. We concentrated on four overlapping mechanisms or catalysts that could galvanize change: examination crises; media reporting; political involvement; and sociocultural

drift. Lastly, we looked into why standard setting systems rarely met the conditions for paradigm shift, concluding that educational assessment systems seem to resist change, and even when they do alter their standard setting processes, these alterations can be either short lived or unsuccessful and therefore abandoned.

References

Baird, J. and Lee-Kelley, L. (2009) 'The dearth of managerialism in implementation of national examinations policy'. *Journal of Education Policy*, 24 (1), 55–81.

Baird, J., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T. and Daugherty, R. (2011) *Policy Effects of PISA*. Oxford: Oxford University Centre for Educational Assessment. Online. https://goo.gl/WKNwfs (accessed 19 June 2018).

Baird, J., Johnson, S., Hopfenbeck, T.N., Isaacs, T., Sprague, T., Stobart, G. and Yu, G. (2016) 'On the supranational spell of PISA in policy'. *Educational Research*, 58 (2), 121–38.

Baird, J. and Gray, L. (2016) 'The meaning of curriculum-related examination standards in Scotland and England: A home–international comparison'. *Oxford Review of Education*, 42 (3), 266–84.

Ball, S.J., Maguire, M. and Braun, A. (2012) *How Schools Do Policy: Policy enactments in secondary schools*. London: Routledge.

Billington, L. (2006) 'Media coverage of examination results, public perceptions, and the role of the education profession'. Manchester: AQA Centre for Education Research and Practice. Online. https://goo.gl/R7cV5G (accessed 19 June 2018).

Black, P., Harrison, C., Hodgen, J., Marshall, B. and Serret, N. (2011) 'Can teachers' summative assessments produce dependable results and also enhance classroom learning?'. *Assessment in Education: Principles, Policy & Practice*, 18 (4), 451–69.

Boereboom, J. (2016) 'University entrance: Always a bridesmaid'. *Education Review*. Online. https://goo.gl/ZbNryy (accessed 19 June 2018).

Botha, R.J. (2002) 'Outcomes-based education and educational reform in South Africa'. *International Journal of Leadership in Education*, 5 (4), 361–71.

Bourdieu, P. (1990) *The Logic of Practice*. Cambridge: Polity Press.

Bruner, J. (1960) *The Process of Education*. Cambridge, MA: Harvard University Press.

Bruner, J. (1996) *The Culture of Education*. Cambridge, MA: Harvard University Press.

Chankseliani, M. (2013) 'Higher education access in post-Soviet Georgia'. In Meyer, H.-D., St. John, E.P., Chankseliani, M. and Uribe, L. (eds) *Fairness in Access to Higher Education in a Global Perspective*. Rotterdam: Sense Publishers, 171–87.

Cross, M., Mungadi, R. and Rouhani, S. (2002) 'From policy to practice: Curriculum reform in South African education'. *Comparative Education*, 38 (2), 171–87.

Dewey, J. (1938) *Experience and Education*. Kappa Delta Pi Lecture Series No. 10. New York: The Macmillan Company.

Erickson, G. (2017) 'Experiences with standards and criteria in Sweden'. In Blömeke, S. and Gustafsson, J.E. (eds) *Standard Setting in Education: The Nordic countries in an international perspective*. Cham: Springer, 123–42.

Fullan, M. (2005) 'The meaning of educational change: A quarter of a century of learning'. In Lieberman, A. (ed.) *The Roots of Educational Change*. Dordrecht: Springer, 202–16.

Fullan, M. (2006) *Change Theory: A force for school improvement*. Centre for Strategic Education Seminar Series Paper No. 157. Melbourne: Centre for Strategic Education. Online. https://goo.gl/J8s1uD (accessed 19 June 2018).

Fullan, M. (2016) *The New Meaning of Educational Change*. 5th ed. Abingdon: Routledge.

Gardner, J., Harlen, W., Hayward, L. and Stobart, G. (2010) *Developing Teacher Assessment*. Maidenhead: Open 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), 245–70.

Herman, J.L. and Baker, E.L. (2009) 'Assessment policy: Making sense of the babel'. In Sykes, G., Schneider, B. and Plank D.N. (eds) *Handbook of Education Policy Research*. New York: Routledge, 176–90.

Husen, T. (1967) *International Study of Achievement in Mathematics: A comparison of twelve countries*. Vol. II. Stockholm: Almqvist and Wiksell.

Jansen, J.D. (1998) 'Curriculum reform in South Africa: A critical analysis of outcomes-based education'. *Cambridge Journal of Education*, 28 (3), 321–31.

Jansen, J.D. (2002) 'Political symbolism as policy craft: Explaining non-reform in South African education after apartheid'. *Journal of Education Policy*, 17 (2), 199–215.

Johnston, K. (2015) 'Top school drops Cambridge exams'. *New Zealand Herald*, 18 September. Online. https://goo.gl/jJc9PX (accessed 19 June 2018).

Kanjee, A. and Sayed, Y. (2013) 'Assessment policy in post-apartheid South Africa: Challenges for improving education quality and learning'. *Assessment in Education: Principles, Policy & Practice*, 20 (4), 442–69.

Klenowski, V., (2009) *Raising the Stakes: The challenges for teacher assessment*. Online. https://eprints.qut.edu.au/43916/2/43916.pdf (accessed 19 June 2018).

Kuhn, T.S. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lee, H., Lee, G. and Openshaw, R. (2013) 'Radical assessment and qualification reform in New Zealand: The rocky road to National Certificate of Educational Achievement'. *World Studies in Education*, 14 (2), 25–45.

Lines, D. (2000) 'A disaster waiting to happen'. *Times Educational Supplement*, 7 April.

Mansell, W. (2013) 'Misleading the public understanding of assessment: Wilful or wrongful interpretation by government and media'. *Oxford Review of Education*, 39 (1), 128–38.

McCaig, C. (2003) 'School exams: Leavers in panic'. *Parliamentary Affairs*, 56 (3), 471–89.

MacCann, R.G. and Stanley, G. (2010) 'Classification consistency when scores are converted to grades: Examination marks versus moderated school assessments'. *Assessment in Education: Principles, Policy & Practice*, 17 (3), 255–72.

Martin, D. (2005) *Report on the 2004 Scholarship to the Deputy Commissioner*. Auckland: State Services Commission. Online. http://www.ssc.govt.nz/upload/downloadable_files/Report-2004-Scholarship.pdf (accessed 27 July 2018).

Meyer, H.-D. and Benavot, A. (2013) *PISA, Power and Policy*. Didcot: Symposium Books.

Morgan, C. and Shahjahan, R.A. (2014) 'The legitimation of OECD's global educational governance: Examining PISA and AHELO test production'. *Comparative Education*, 50 (2), 192–205.

Murphy, R. (2013) 'Media roles in influencing the public understanding of educational assessment issues'. *Oxford Review of Education*, 39 (1), 139–50.

Newton, P.E. (2005) 'Threats to the professional understanding of assessment error'. *Journal of Education Policy*, 20 (4), 457–83.

Philips, D. (2006) 'The contribution of research to review of national qualifications policy: The case of the National Certificate of Educational Achievement (NCEA)'. Paper presented at the British Educational Research Association Annual Conference, University of Warwick, 6–9 September. Online. www.leeds.ac.uk/educol/documents/157310.htm (accessed 19 June 2018).

Powlesland, B. (2017) *What PISA tells us about grade inflation in NCEA*. Online. https://goo.gl/HMxD14 (accessed 19 June 2018).

Rhoades, K. and Madaus, G. (2003) *Errors in Standardized Tests: A systemic problem*. National Board on Educational Testing and Public Policy. Boston: Boston College. Online. https://files.eric.ed.gov/fulltext/ED479797.pdf (accessed 19 June 2018).

Sayed, Y. and Ahmed, R. (2011) 'Education quality in post-apartheid South African policy: Balancing equity, diversity, rights and participation'. *Comparative Education*, 47 (1), 103–18.

Schmidt, M.J. (2017) 'The perils of outcomes-based education in fostering South African educational transformation'. *Open Journal of Political Science*, 7 (3), 368–79.

Scott, D. (2011) 'Assessment reform: High-stakes testing and knowing the contents of other minds'. In Berry, R. and Adamson, B. (eds) *Assessment Reform in Education: Policy & practice*. Dordrecht: Springer, 155–63.

Sirotnik, K.A. (1998) 'Ecological images of change: Limits and possibilities'. In Hargreaves, A., Lieberman, A., Fullan, M. and Hopkins, D. (eds) *International Handbook of Educational Change*. Dordrecht: Kluwer Academic, 181–97.

Sirotnik, K.A. (2005) 'Ecological images of change: Limits and possibilities'. In *The Roots of Educational Change*. Dordrecht: Springer, 169–85.

Stanley, G., MacCann, R., Gardner, J., Reynolds, L. and Wild, I. (2009) *Review of Teacher Assessment: Evidence of what works best and issues for development*. Report commissioned by QCA. Oxford: Oxford University Centre for Educational Assessment. Online. https://goo.gl/YQH39d (accessed 19 June 2018).

SSC (State Services Commission) (2005) *Report on the 2004 Scholarship*. Online. www.ssc.govt.nz/node/4199 (accessed 21 February 2018).

Tomlinson, M. (2002) *Inquiry into A Level Standards: Final report*. December 2002. Department for Education and Skills. Online. http://dera.ioe.ac.uk/19112/ (accessed 19 June 2018).

Vallender, G. (2009) 'External examinations beyond national borders: New Zealand and the Cambridge International Examinations'. In Vlaardingerbroek, B. and Taylor, N. (eds) *Secondary School External Examination Systems: Reliability, robustness and resilience*. Amherst, NY: Cambria Press, 291–303.

Chapter 16

Setting standards in national examinations: What we have learnt

Tina Isaacs

In this book we set out to investigate, document, analyse and evaluate setting and maintaining standards in national curriculum-based school leaving or university entrance examinations. Outcomes from these examinations are critical to the life chances of the students who sit them and in many cases to judgements about the schools in which the examinations are sat. School leaving and university entrance examinations in many jurisdictions stand as a proxy for the quality of the education system itself. Through conceptual and case study chapters, this book has explored how standards are defined and how those definitions are enacted, as well as trenchant system issues and challenges. It has put forward three paradigms for understanding educational assessment, provided insights to insider research, and offered a new theoretical conceptualization of the meaning of examination standards using an ecological model.

We found that while education cultures are different in different places, as are standard setting and maintaining policies and practices, a number of themes arose both from the Standard Setting Project itself and from the additional research for this book. One incisive issue was the role of fairness and social justice in educational assessment in general, and in examinations in particular. While wanting to provide the most efficacious and acceptable means of distinguishing among test-takers, a number of jurisdictions, such as South Africa and Chile, recognized – and some struggled with – how to reconcile this differentiation with past and present social and economic injustices.

There was no consensus on the role that teacher judgement and internal assessment could and should play in setting school leaving standards. Continuous assessment was deemed by some to allow a greater emphasis on deep learning and the skills that students would require to be successful in their university programmes and later employment or further study. Abiding concern was expressed about whether – and how – to assess

twenty-first-century skills such as creativity and innovation, critical thinking, problem solving, decision making, meta-cognition, communication, team work and citizenship (Adamson and Darling Hammond, 2015). However, comparability concerns and maintaining standards over time weighed against an over-reliance on teacher judgement, with some jurisdictions such as Sweden and Queensland shifting emphasis or moving back, at least partially, to more customary external testing and examining.

There was a general unease – and yet a preoccupation – with what standards are and whether or not they are rising or falling. Questioning about the efficacy of, for example, the French baccalauréat, the English A levels and the Georgian Unified National Examinations has directly challenged whether or not the 'right' standard is being set for university entrants. In many cases high pass rates have caused political and media scrutiny. This, in turn, has led people to question the system that has produced these results. When a lack of trust in the examinations process takes hold in a system, stakeholders such as students, parents and teachers can lobby for changes or simply become cynical. However, we also found that despite questioning and doubt, most standard setting systems have remained relatively stable, with some tinkering around the edges rather than instituting deep-seated change.

Contributions this book has made

National examination standards are a central currency and of great importance for people's lives around the world. Yet, the research literature has largely ignored the ways in which standards are defined and set in most countries. If one's only understanding were through published work, it would appear that psychometrics is the main way in which examination standards are set. However, for the first time, we have documented that this is not the case. Examination boards around the world do not primarily use a psychometrics paradigm. This book is therefore a useful addition to the literature for depicting the ways in which standards are set in the nine jurisdictions included in the book.

Transparent procedures are not always publicly available in every country, and where they are, they may be written at such a high level that it is difficult to untangle exactly what happens in practice; the weight given to different sources of evidence, the sequence of events, who has the decision-making power and so on. The chapters in this book are far more transparent than previously available documentation and are all presented in English, as language has been another challenge to accessibility.

Widening the range of jurisdictions included in the literature on standard setting has challenged the notion that psychometrics is the only, or dominant, paradigm. Public examination standards are being set for millions of students around the world every year without reference to the psychometrics paradigm. In this book we have introduced two other educational assessment paradigms – curriculum-based and outcomes-based. Commentary and practice relating to these approaches already existed in the literature, but they had not previously been contrasted as separate paradigms, which have different underlying belief systems and procedures. A prevailing view is that the psychometrics paradigm is superior. This position has a long history, as we noted in Chapter 1. The International Examinations Inquiry conducted in the 1930s was designed to promulgate more scientific forms of examining across the Atlantic to Europe (Lawn, 2008). Although that early attempt was a failure, there were some successes and some of the countries involved (e.g. Sweden) were influenced by the Inquiry and have subsequently utilized psychometric techniques to a larger extent than in other countries' national examinations (e.g. England and France). Some have argued that the techniques from curriculum-based and outcomes-based paradigms are weaker forms and that those countries do not use psychometric techniques because they do not have the expertise. Certainly, it is hard for examination boards to recruit technical staff with the training and skills to conduct psychometric analyses. However, this is true for quantitative techniques in education in general and there have been strategic initiatives in a number of countries to try to rectify this, through research funding bodies.

Skill shortages are no doubt part of the answer to the lack of universal uptake of psychometrics, as no policymaker could change the approach overnight due to lack of people to make it happen. This is not the only explanation, but to know that requires both an overview of the field internationally today and a historical analysis of educational assessment. Curriculum-based and outcomes-based techniques are championed in some jurisdictions because they better suit the cultural and historical contexts in which the examinations are situated.

Returning to the superiority of the psychometrics paradigm, we have presented it as one paradigm among three and have not taken the position that any of them is more suitable than the others. Notwithstanding, we have pointed out that in practice, standard setting systems have tried to borrow across paradigms in their procedures. While there are attractions to this, such as using the latest techniques and trying to address criticisms of the current techniques in use, it can cause problems. Coherence of the approach

can be lost as practitioners try to rationalize the links between the questions they are trying to address and the evidence produced through the cross-paradigm techniques they are utilizing. Language used often creeps across paradigms, which can then mean that the utterances are counter-cultural, with all of the problems that can bring. Ultimately, there is no truth to find in the process of setting national examination standards, and the job is about marshalling the evidence in as rigorous a manner as possible. What counts as rigorous depends upon underlying views about what you are trying to assess and how that is best achieved. These views and aims are distinctively different across the paradigms.

We claim that the standard setting methods depicted in this book are an addition to the literature in itself. Oftentimes audiences will ask at the end of a presentation on the setting of examination standards, 'Yes, but is that what is really done, or is it just what is claimed publicly?' This leads to a range of questions about the authenticity of the descriptions in this book. Why should they be trusted? As we ourselves say in Chapter 1, standard setting is highly political; that being so, are the case study authors free to write exactly how things are done? After all, we selected examination board insiders to participate in this project. For the first time in the literature, we have tackled this issue. The project has led to a set of guidelines for examination board insiders to use in navigating research projects. It is based upon the literature on insider and action research, elite interviewing and on theories of researcher positioning. Our project was therefore reflective about these issues and, as such, we make positional claims rather than truth claims. It is for the reader to decide whether the text is trustworthy, but the authors' job is to be transparent enough about who they are and how their claims can be evidenced to allow the readers to draw their conclusions. Certainly, it can be easier for insiders to discuss standard setting in historical perspective because the implications for the people involved are likely to have changed over time.

Examination policy can be fast-changing in some of these contexts too, so the material here represents a particular snapshot in time. 'Historical practices' is the highest level of the ecological model of definitions of examination standards presented in Chapter 14. The ecological model is also an addition to the literature, as it explains why so many definitions have coexisted and why examination boards are not ready to simply disregard challenges arising from a range of definitions. Our investigation of examination standards definitions also showed that some definitions cross the levels of the ecological model. These two definitions – attainment-referencing and standards-referencing – are indistinguishable conceptually,

but they have sprung from different paradigms; curriculum-based and psychometrics respectively. Using our ecological model and the paradigms, we have for the first time produced a framework for classifying national examination standards definitions in use in different jurisdictions.

Most standard setting methods involve the integration of information from a range of sources. Policy descriptions typically list all of these sources. How the information is integrated and the weight given to these different sources of information can be difficult to comprehend. Use of the range of information available may even be purposefully under-determined in the policy context to allow for adaption to different circumstances by the standard setting decision makers. A body of methodological work has conceptualized the integration of mixed methods research data (e.g. Teddlie and Tashakkori, 2016). Here, we have applied those techniques to describe how the standard setting methods used in the jurisdictions in the book claim to weight the qualitative and quantitative evidence and the order in which this occurs. Conceptualizing the standard setting process as a mixed methods study is also new to the literature and helps to better explain the integration of the wide range of information available in the processes. Showing standard setting processes in this way is also a higher level of formalization of the procedures.

From our experience of the field, we expected that each jurisdiction's standard setting processes would be strongly affected by its educational, social and political cultures and that standards would be contested, even in the least transparent regimes. That hypothesis proved correct, but despite questioning, disputation and controversy, standard setting procedures proved remarkably resistant to fundamental change. Almost nowhere did we encounter a radical rethinking about standard setting, despite large-scale 'crises' such as that in England in 2002. Instead, jurisdictions made accommodations to existing practices – in England, for example, shifting the balance in awarding procedures away from emphasis upon examiner judgement to stronger reliance on statistical predictions through comparable outcomes procedures.

Using a framework inspired by Thomas Kuhn's paradigms in *The Structure of Scientific Revolutions* (1962) as well as Michael Fullan's theoretical models for educational change, we grappled with the dilemma of why, despite dissatisfaction among some stakeholders and public opprobrium, sometimes on a large scale, the education establishment was unable or unwilling to promote change. Political and policy considerations dictated cautious approaches. Whereas in the physical sciences it is possible to find examples where ideas posited by the scientific community prevailed

over political opinion (although climate change scientists might dispute this contention), in educational assessment politicians who are responsible for education policy can, and do, use standard setting as well as wider curriculum and assessment issues as a means of reinforcing political agendas that rarely coincide with challenging the status quo. The short lifespan of many governments or education ministers also works against radical change. While politicians might want to point to advances in educational outcomes during their tenure, often through 'raising standards', those tenures tend to be too short term for root and branch reform, even if education-based stakeholders and the public were exerting pressure for it. Decision making therefore can be a matter of expediency and political necessity.

The case studies

It is extremely challenging to summarize briefly the contributions made by each of the jurisdictions that provided case studies for the book. What follows attempts to highlight, in alphabetical order, some of the challenges of successfully setting and maintaining standards in school-leaving examinations, drawing out additional issues and cross-cutting themes.

Fairness and equity are themes that recurred in some of the case studies. Chile now has had 14 years' experience with its university entrance tests, the *Prueba de Selección Universitaria* (PSU), and against hopes and expectations, the outcomes gap between students from low and high socio-economic backgrounds has increased rather than shrunk. Universities are autonomous and set their own entrance criteria, and only the most selective institutions rely on PSU. Its advent has also increased the power of the National Council of Rectors of Chilean Universities (CRUCH), who support the PSU despite its drawbacks. But resistance to unquestioning reliance on the PSU is growing, and calls have been made to diversify the testing programme, especially for those students who have followed more technical and vocational pathways. Some also question the validity of the PSU tests, claiming that by focusing solely on content knowledge, students are not being tested on the skills they will need to be successful in university.

Over the past decade the Office for Qualifications and Examinations Regulation (Ofqual) has introduced a system of comparable outcomes in England's two secondary school curriculum-based qualifications, GCSEs and A levels, which has stabilized standards maintenance and slowed down grade inflation. While both statistics and human judgement come into play, it is the former that is more heavily paid heed to. This does not prevent close scrutiny by stakeholders, especially the media, of the outcomes of standard setting and maintaining each year. One consistent refrain had been

that the examinations were being 'dumbed-down' and were therefore less useful tools in the university selection process than they could be, and this idea also reinforced the idea that schools and further education colleges were somehow 'getting away' with providing a less rigorous education. With the advent of comparable outcomes, standards-related questions are being asked. Is this newer approach preventing real improvements in student performance from being recognized? Due to the complexity and political underpinnings, as well as the very public and transparent nature of England's examination system, challenges to it are likely to continue, regardless of which political party is in power.

The practices and procedures for setting standards in the French baccalauréat are less transparent than in many other countries. Around 80 per cent of 18-year-olds sit some form of the baccalauréat and the pass rate is reasonably high – 88.5 per cent in 2016. There seems to be less questioning of the examinations' standards – and standard setting processes – in France than elsewhere. This may be rooted in the fact that the national examination and university entrance system masks a separate, more elitist system, found in the *grandes écoles*. Students wanting to attend the *grandes écoles* are chosen after two years of highly competitive preparatory courses; admission does not rely on baccalauréat outcomes although applicants are asked to obtain one. Those who pass the baccalauréat are eligible to attend any public university in the subject of their choosing, but not to enter the elite further training courses. However, students' failure rate in the first two years of higher education is high, which has caused some consternation (Bodin and Orange, 2017). The discrimination function that is found in other standard setting for school leaving examinations seems to be far weaker in France than elsewhere. (In January 2018 the government announced reforms to the baccalauréat that addressed some of the criticism presented above. Changes to be in place in 2021 will allow students to choose between more in-depth *spécialités* and have been made to ensure that both curriculum and examinations better prepare students for the requirements of higher education. Examination results will be included in applications for both types of higher education.)

Post-Soviet Georgia still struggles with standard setting and maintaining in an atmosphere in which fighting corruption in university admissions (and within the university system itself) is paramount. In the past wealthy and influential people were able to buy their children's way into higher education. Now, however, university entrance examinations (Unified Admissions Examinations, or UNE) have cut scores that are little above what an applicant might receive by guesswork. Universities accept

students on their programmes who are not university ready, in part because universities are dependant for their survival upon students' tuition fees. School accountability measures do not include student performance. In the face of these challenges, there is little incentive to introduce rigorous and discriminating standard setting procedures or to make the examination system transparent to stakeholders. UNE are both valued and trusted by the teaching profession and the public at large, perhaps more a testament to what they replaced than a vote of confidence in the examinations themselves.

Ireland's school leaving certificate and its examination system date from the early twentieth century and Ireland enjoys a large tertiary participation rate. Centrally administered university admissions based on set entry criteria and rank ordering points-based methods play a role in system stability, although this does lead to increasingly specialized university course offerings as higher education institutions (HEI) compete for the best students. Comparability of standards across subjects is assumed and all are weighted equally for university entry (with the exception of advanced mathematics). Standard setting through post hoc changes within the marking process is a particular feature differentiating Ireland from most of our other case studies. Grade boundaries are fixed and adhere to a predetermined percentage of available marks. If during the marking process it transpires that 'standards over time' will not be preserved, the mark schemes are altered to achieve more acceptable grade distributions. Of concern among the education community is that this standard setting method can stifle innovation, since in order to achieve year on year stability the examinations must be somewhat predictable.

Having had an externally moderated school-based assessment and university entrance system since the 1970s (complemented by a core skills test), Queensland has very recently decided to introduce externally set components worth 25 per cent of students' overall attainment rating. This will bring it more in line with Australia's other states and territories, which employ a combination of external and school-based assessments wherein external examination results are used to scale school-based ones. Starting in 2019, moderated school-based assessment will contribute 75 per cent toward a student's subject result except in mathematics and science, where it will be 50 per cent External components, however, will not be used to scale internal components, as in the rest of Australia. Students' rank ordering for university entrance will be based on inter-subject scaling of their best five subjects. While there is continued support for, and a high weighting of, teacher-led summative assessment, critics argued that the moderation system was not sufficiently robust, reliable, or fair, especially in

mathematics and the sciences. Queensland's decision to include externally set and marked assessments into the overall mix brings it more in line not only with other Australian states and territories but with much of the rest of school leaving assessments internationally.

Apartheid's legacy weighs heavily on the South African examination system and its National Senior Certificate (NSC). Attempting to overcome regional, class and racial differences, since 2008 students in the 12th grade have sat the same externally set school leaving examinations, which count for 75 per cent of the examination outcome; the other 25 per cent is moderated school-based assessment. The NSC's natural technical teething problems are exacerbated by policies put in place to ensure fairness and social justice. Different examination levels – higher and standard – were abandoned, producing an unusually high proportion of students who were deemed to be university ready. Universities have reacted with scepticism and expressed doubts that the NSC is a good predictor of later academic performance in some subject areas. They argue that applicants are not prepared for university work and therefore drop out of their university programmes, something we have already noted happens in France. And as Howie points out in her commentary, another fallout from apartheid is the lack of technical capacity in setting, moderating and marking examinations as well as in quality assurance processes.

Like Queensland, Sweden relies on school-based assessment to determine students' eligibility for higher education. And also like Queensland, Sweden is once again grappling with the role of externally set examinations, the national tests. These tests were reintroduced to complement teacher judgement and to help make those judgements more reliable, and in subjects where such tests are available, teachers give them a great deal of weight. If teachers' judgements and national test outcomes diverge too much, this is seen as a serious problem. Issues of differences in standard setting between schools and over time have also arisen. Some have argued for a stronger reliance on the national tests, contending that they are more reliable and fair. They can also impede grade inflation. Currently the national tests are not high stakes, but that may change. Discussions are ongoing about using national test outcomes to make comparisons over time or to evaluate school performance; a framework outlining test development, interpretation and use is being worked on.

Alone among our case studies, the United States has no national curriculum and therefore no curriculum-based school leaving examinations. Some states have their own tests, such as the New York State Regents examinations, but the only national tests for upper secondary students are

the SATs and ACTs, which it can be argued are curriculum-aligned to a certain extent, but not curriculum-based because they are divorced from national programmes of study. That leaves the Advanced Placement (AP) tests in an almost unique position – they have curricula that are available nationally and a set of examinations attached to those curricula. In the absence of nationally agreed curriculum standards, AP has become the *de facto* assessment system for higher achievers, since they are meant to be at first year university standard. While this means that an increasing number of schools and students are accessing AP, students of average and lower achievement are left without valid and reliable curriculum-based reflections of their high school performance, which puts a large burden on all but the most selective universities in making admissions choices.

The limitations of our study

We faced a number of difficult decisions in research design for the Standard Setting Project. We acknowledged our positions as insider-outsiders, and the benefits and limitations of such positioning. Starting our reflective processes from our own positions, we concluded that senior examination board personnel would be best placed to access and share the detailed knowledge of policies, processes and approaches that we were seeking. Using such insiders necessarily brought limitations, which we aimed to offset through careful research design and appropriate support for project participants.

Our aim was to present a range of contrasting cases to increase knowledge of standard setting practices around the world. Finding the key people to participate was sometimes a challenge, since we wanted to work with those who were intimately involved in the standard setting process. Our approaches directly and through our networks failed in a number of countries. Unfortunately, the political pressures that we have described also affected some of our potential participants, and a small number found that they could not secure organizational or policy approval to take part. One or two of these are currently in the midst of major qualifications reform. Despite this, we secured 12 project participants from a range of jurisdictions around the world, and our purposive sampling of cases ensured that the project depicted a variety of approaches to standard setting, with different assessment formats and use of differential cut scores, as well as wide geographical spread and cultural distinctiveness, including cases from developed and developing systems, and systems currently subject to far-reaching reform of qualifications. We were struck by the fact that many of the symposium participants, who were best placed to discuss standard setting within their jurisdictions, were eager to learn about what happens

in other jurisdictions. Some of these policy implementers did not seem to have participated in the sort of networks that are widely used by senior examination board colleagues, academics and policymakers such as the International Association of Educational Assessment (IAEA) and the Association for Educational Assessment – Europe (AEA-Europe). We are hopeful that our work is a first step to creating such networks, and there is already evidence of collaboration and communication outside the project.

We also carefully considered sources of evidence. As well as highlighting the issues inherent in insider researcher, our pilot study on Scotland and England showed us that documentary and archive evidence is not enough on its own (Baird and Gray, 2016). We knew therefore that we needed to use more than one source of evidence. We considered in-depth interviews of wider participants and direct observation as possible additional sources, but given the international nature of the project, these two sources of evidence would have required people and budgetary resources beyond our means. Instead, we provided alternative perspectives and rival explanations of the phenomena presented. Participants agreed at the outset that their own position would be contrasted with those of two other in-country experts, and that while we would share with them the views of those experts, they would have no right of reply except to correct factual inaccuracies. For each of the country cases in this book, readers are able to read the insider's account and to compare this with the two commentaries provided. We also carried out a series of interviews with our insider researchers in which we sought to challenge and confirm participant accounts of their own systems. Our thematic chapters draw on this interview data to provide commentary on the accounts presented by our insider researchers. The interviews challenged our own understandings of standard setting and resulted, we hope, in the co-creation of knowledge.

Readers of this book can survey the range of evidence sources and make their own judgements on the authenticity and trustworthiness of the data presented. We claim that the cases captured in this book present a characteristic array of systems from around the world, and that the range of data sources represents a strenuous attempt to depict a variety of positions and to make transparent what was previously opaque. We note again, though, that this research is positional and leave the reader to make their own judgement on the positions presented.

Future research

This project has shed a great deal of light on the ways in which standards are defined and set in the range of jurisdictions included in this book and

beyond. Notwithstanding, not all aspects of the project can be reported here, and there are some clear next steps for research that arise from the project and the literature. We have codified to some extent the approaches to standard setting that were taken, but there is clearly more that could be done to depict and classify the different approaches. Extending the work on mixed methods to a fine level of detail in a range of countries would be a very useful next step and one that would allow better comparisons to be made. And while we defined standard setting in Chapter 4 as 'any process by which raw marks are converted into the reported outcome', further research is needed around this broader, less technocratic concept, how the definition might be applied across the different paradigms, and where in the ecological model standards are set within education and assessment systems.

Additionally, more nuanced work on the commonalities and differences between assessment paradigms is required. Their applications to case studies in practice would be illuminating.

As part of the Symposium held at Brasenose College in Oxford in March 2017, representatives from different countries were asked to produce an outline of the organizational structures involved in national examinations and the responsibilities each of those held. It became apparent rapidly that this technique was very useful in showing how the culture and context of examinations differed and was represented in the institutional arrangements. More formal documentation of this kind would be an original introduction to the literature. Insights could be gained by further work on standard setting processes, procedures and policies in the three jurisdictions whose representatives attended the symposium but which do not have case studies in this book: Hong Kong, South Korea and Victoria. For example, Hong Kong has attempted to put in place curriculum-based assessment models over the last 30 years, some of which, like the Target-Oriented Curriculum of the 1990s, have been abandoned (Carless, 2012). More recently, school-based teacher-graded assessments have been introduced that contribute between 15 and 25 per cent of the subject marks of the Hong Kong Diploma of Secondary Education (HKDSE). Whether these reforms represent paradigm shifts or accommodation within the existing systems could be usefully explored.

In Chapter 1 we discussed the Americanizing influences that can be observed in some systems. Post-colonial inheritances in assessment systems also could be usefully explored further. England, France and the US have influenced strongly the examination and standard setting legacy of those nations and jurisdictions within their former – and current – spheres of influence. One only has to look at the impact that England has had on

commonwealth countries to realize that the English examination system has spread massively. Equally, the effects of other colonial empires upon the education systems and their standards have not been part of the analytical approach in the book, but might usefully be so in future work.

Finally, Foucault (1977) argued that assessment was part of the power of normalization in everyday society. Through assessment we are controlled and even internalize the judgements, thereby normalizing our own behaviour. Thus, questions about who decides what is to be valued in assessments and to define their standards are as profound as they are insidious. In this book we barely touch upon the power dynamics underlying educational assessment, but a sociological analysis would foreground them, seeing them as the most important feature of this area of study. Due to our own positions and the expertise that we bring, we have prioritized other aspects of the research, but we recognize that whosoever has the power to decide examination standards in policy and in practice is a pressing area for research. Contrasting those arrangements across jurisdictions in a comparative analysis will also be a valuable contribution to understanding examination standards and their relations with the societies they serve.

References

Adamson, F. and Darling-Hammond, L. (2015) 'Policy pathways for twenty-first century skills'. In Griffin, P. and Care, E. (eds) *Assessment and Teaching of 21st Century Skills: Methods and approach*. Dordrecht: Springer.

Baird, J. and Gray, L. (2016) 'The meaning of curriculum-related examination standards in Scotland and England: A home–international comparison'. *Oxford Review of Education*, 42 (3), 266–84.

Bodin, R. and Orange, S. (2018) 'Access and retention in French higher education: Student drop-out as a form of regulation'. *British Journal of Sociology of Education*, 39 (1), 126–43.

Carless, D. (2011) *From Testing to Productive Student Learning: Implementing formative assessment in Confucian-heritage settings*. New York: Routledge.

Foucault, M. (1975) *Discipline and Punish: The birth of the prison*. New York: Random House. Online. https://zulfahmed.les.wordpress.com/2013/12/disciplineandpunish.pdf (accessed 20 July 2018).

Kuhn, T.S. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Teddlie, C. and Tashakkori, A. (2016) *Foundations of Mixed Methods Research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks, CA: SAGE Publications.

Appendix A

Attendees at the Brasenose College Symposium, 28 to 30 March 2017, Oxford University

Jo-Anne Baird	Oxford University Department of Education, UK
Louise Benson	National Foundation for Educational Research (NFER), UK
Beth Black	Office of Qualifications and Examinations Regulation (Ofqual), UK
Tom Bramley	Cambridge Assessment, UK
Rob Coe	Centre for Evaluation and Monitoring, Durham University, UK
Sally Collier	Office of Qualifications and Examinations Regulation (Ofqual), UK
Wenjun (Elyse) Ding	Oxford University Department of Education, UK
Barbara Donahue	Standards and Testing Agency (STA), UK
Roger-François Gauthier	University Paris Descartes, France
Kristine Gorgen	Oxford University Department of Education, UK
Lena Gray	Assessment and Qualifications Alliance (AQA), UK
Therese N. Hopfenbeck	Oxford University Department of Education, UK
Tina Isaacs	UCL Institute of Education, UK
Kate Johnson	Standards and Testing Agency (STA), UK
Ben Jones	Assessment and Qualifications Alliance (AQA), UK
Kate Kelly	Assessment and Qualifications Alliance (AQA), UK
Iasonas Lamprianou	University of Cyprus, Cyprus

Yong-Sang Lee	Korea Institute for Curriculum and Evaluation (KICE), South Korea
Anna Lind Pantzare	Umeå University, Sweden
Joshua McGrane	Oxford University Department of Education, UK
Hugh McManus	State Examinations Commission, Ireland
Michelle Meadows	Office of Qualifications and Examinations Regulation (Ofqual), UK
Deanna L. Morgan	The College Board, USA
Paul Newton	Office of Qualifications and Examinations Regulation (Ofqual), UK
Diana Ng	Oxford University Department of Education, UK
Dennis Opposs	Office of Qualifications and Examinations Regulation (Ofqual), UK
Alejandra Osses	University of Santiago, Chile
Gareth Pierce	Welsh Joint Education Committee (WJEC), UK
Nicky Platt	UCL IOE Press, UK
Alastair Pollitt	Cambridge Exam Research, UK
Derek Richardson	Pearson, UK
Mary Richardson	UCL Institute of Education, UK
Alex Scharaschkin	Assessment and Qualifications Alliance (AQA), UK
Emmanuel Sibanda	Umalusi, South Africa
Gordon Stobart	UCL Institute of Education, UK
Rachel Taylor	Office of Qualifications and Examinations Regulation (Ofqual), UK
Chong Sze Tong	Hong Kong Examinations and Assessment Authority, Hong Kong
Natalie Usher	Oxford University Department of Education, UK
Marieke van Onna	CITO, Netherlands
María Leonor Varas	University of Santiago, Chile
Anna Kristina Wikström	Umeå University, Sweden
Guoxing Yu	University of Bristol, UK
Nadir Zanini	Office of Qualifications and Examinations Regulation (Ofqual), UK
Nathan Zoanetti	Victorian Curriculum and Assessment Authority (VCAA), Australia

Appendix B

Guidelines for the exam board insider researcher

Lena Gray

Many guidance documents provide advice and checklists on how to carry out research projects, and many have useful things to say about aspects of qualitative research, action research and insider research (for example, Denscombe, 2010; Bell, 2005; Blaxter *et al*., 2006). In exam board research, we need to consider some key points from descriptions of several different research methods, and a distillation of those into one document may prove helpful. This section sets down some lessons and pointers that I have found useful in carrying out and reflecting on my own insider research. It has been compiled following a search of existing guideline documents – although given the size of the field, not one that claims to have involved comprehensive searching or systematic review.

The suggestions given have been tested through knowledge exchange with insider researchers, and their views were sought on how useful and practical they found previous drafts of these guidelines.

The advice below is not intended as a guide to research methods or ethics. It is assumed that it is addressed to researchers who already have established practices, and who wish to reflect on how better to create conditions to ask the question, 'What are we doing here?' Its focus is purely on how to open up spaces that facilitate openness and transparency, and allow the insider researcher to 'speak truth to power' (American Friends Service Committee, 1955). As such, it is intended to supplement existing knowledge and ways of working and does not cover all aspects of research design and planning, data gathering and analysis and reporting.

The guidelines are arranged around the four stages of research suggested by Costley *et al*. (2010 – see below).

For each stage of the research process, some general guidance notes are provided. For the first three stages, these are followed by a text box containing a checklist of possible actions and/or key questions to ask. The checklists are adapted from existing guidelines and checklists on conducting research – mainly those on insider research (Zeni, 1998; Coghlan and Brannick, 2010; Costley *et al*., 2010; BPS, 2014; Kemmis *et al*., 2014).

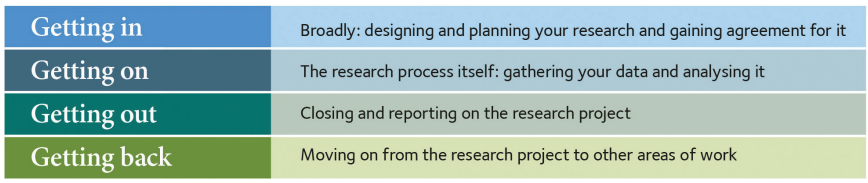


Figure B.1: The four stages of research suggested by Costley *et al.* (2010, Chapter 5)

Getting in

The first task in any research project is to define the nature of the research. There is much advice on ways to go about this, but Coghlan and Brannick point out some particular problems for the insider researcher. For example, it may be tempting to think that senior colleagues may be 'won over' to the need for the research if they are presented with it as a way to solve a problem. Coghlan and Brannick advise against that approach: 'It may be that organisational members embrace problems with a sense of loss, wondering about the organisation's ability to reach a satisfactory resolution' (Coghlan and Brannick, 2010: 54). As already touched upon in the discussion paper, for exam board researchers working in organizations arguably already subject to risk avoidance and scapegoating, talk of 'problems' may not be the best way to convince colleagues that your research will be helpful. On the other hand, they argue, framing your research in terms of opportunities may also be less than helpful, engendering excitement, encouraging divergent thinking and creating a risk-taking culture around the project. Better, they suggest, to frame the project in terms of 'issues', which they view as a neutral term (Coghlan and Brannick, 2010: 54). However, a glance at synonyms for 'issues' in any thesaurus might suggest the opposite, and perhaps the best advice is to think carefully about your language: 'topics' might be a more neutral term, or 'questions'.

Even if you are successful in framing your project in a neutral way, colleagues – including senior colleagues – may have a range of concerns about the work. Some of these may be purely practical: for example, concern might be expressed over the amount of time or resources that will be involved. In effect, they are giving you time to carry out this research, and they want to be assured that the work will produce benefits for the organization.

Most frameworks for insider research emphasize issues around consent to carry out the research. This guidance appears to make an implicit

assumption that the researchers are working at other roles and need to make themselves known as researchers. Providing assurances of confidentiality for colleagues is seen as a key part of this process. Ethical frameworks almost always stress the need to avoid deception of research subjects, but if you are an insider researcher whose colleagues know that you are researching your own workplace's practices, then this becomes more complex. If research is, in fact, your day-to-day job, while in one way you are always open with colleagues, in another you are constantly in danger of practising deception: are your colleagues always aware of the particular work you are doing – its aims and purposes?

In the context of insider research, 'getting in' is less about negotiating access, consent and confidentiality, and more about some difficult upfront conversations about the possible short- and long-term ramifications of carrying out and sharing the research. The British Psychological Society's (BPS) *Code of Human Research Ethics* (BPS, 2014) reminds us that scientific integrity requires clear aims: 'It is important that the aims of the research are as transparent as possible to ensure that it is clear what the research intends to achieve' (BPS, 2014: 10). Whether you are proposing the research topic, or someone else is proposing it to you, extended negotiations may be necessary to achieve this transparency. These negotiations should include overcoming concerns and highlighting benefits: in effect, you will have to sell your research to colleagues.

In the complex political world in which exam boards operate, public trust is both essential and fragile, and research always carries risk. Research within exam boards is likely to fall into one or more of the categories defined by the BPS as 'more than minimal risk', including research involving access to confidential information; research involving access to potentially sensitive data; and research that may have an adverse impact on employment or social standing (for example, discussion of an employer, or discussion of commercially sensitive information). Importantly, too, for the exam board researcher, the BPS guidelines conclude that: 'Risk analysis should not only be confined to considering the interests of the primary participants, but should also consider the interests of any other stakeholders' (BPS, 2014: 13–14).

When you are proposing or developing the research topic, risk assessments and negotiations around it are essential. For exam board researchers assigned a research project by superiors, it can be tempting to assume that such considerations do not apply – but senior personnel may not have research experience, and will not have time to think through a

proposal in the same amount of detail as you do. If you do not want to find yourself in the frustrating position of having invested time and effort in a research project only to have senior colleagues ask for it to be stopped at a later stage, then you need to try to anticipate as many of the risks and issues as possible, and discuss these upfront with key decision makers. The organization's hierarchies and decision-making structures will be important here, and it will be useful to you if there are explicit and agreed responsibilities for signing off research proposals, research outputs and research dissemination strategies. It is important that you are absolutely clear about which individuals or groups have this responsibility and, if there are different individuals or groups involved, you should spend some time reflecting on how these might interact with each other.

How to prepare the way to ensure the best chance of success for your insider research project

As well as your usual approaches to research design and planning, you should consider taking some or all of the following steps before your research begins.

Achieving buy-in

- Think carefully about how you will frame and describe your research. Consider talking about research topics or questions rather than problems, issues, or even opportunities.
- Establish credibility – just because you are part of an organisation and have support or high-level agreement for your project, you should not assume that everyone in the organisation will see the value of your project. Some research personnel work within an organisation, but slightly detached from it; at a personal level, making sure that your colleagues know you and your work can be really helpful when you need to discuss specific research projects with them.
- 'What's in it for me?' You may find it useful to ask senior colleagues and/or research participants to define what they would like to get out of the research.
- Describe the purpose of your research and its benefits for your colleagues and/or the organisation as clearly as possible: sell your research but do not overstate the benefits.
- Engage with colleagues to find out their concerns and discuss how these will be overcome.
- Consider using a risk analysis tool to document all of the possible risks and mitigating actions.

Ethical considerations

- Before beginning the work, you will have to take extra steps to make sure that colleagues (including colleagues more senior to you) are aware that the findings of the research cannot be guaranteed to be positive, or as expected.
- You will need to clarify for colleagues what use you will make of normally confidential information. Usually, what you will be doing here is reassuring colleagues that confidential information will not be made public. Making sure that colleagues know your research code of practice and have faith in how you implement it may be a longer-term task that is necessary to underpin trust in particular research projects.
- Work to be shared – even internally – will never be completely anonymous. You should take time to ensure that everyone involved or affected is aware of this. If you are planning to share your work outside the organisation, you also need to make sure that relevant senior staff are aware of possible ramifications for stakeholders or customers.
- Ask yourself what negative or embarrassing data you can anticipate emerging from this research. Might the organisation or individual colleagues be harmed (reputationally, professionally, financially)? Discuss the risks with these people and set out for them the precautions you will take to protect individuals, teams and/or the organisation.

Figure B.2: Getting in

Many social science codes of research practice emphasize causing no unnecessary harm, and this may be complex for the exam board insider researcher. We saw earlier that there is a range of stakeholders who may

have an interest in your research; some of these may be directly impacted by it, and you may have to make difficult choices, balancing benefit and harm to different groups. Again, the key is to be explicit and to make sure that the relevant decision-making individuals and groups are aware that outputs may benefit some colleagues or stakeholders but harm others. Do not forget, too, about senior stakeholders: while it may be difficult to imagine senior colleagues as vulnerable, in terms of publication of reports about aspects of organizational activity, it is senior staff who will bear responsibility and whose lives may be affected by your research. As an employee, you have a right to expect them to do you no harm, but as a researcher, you have an ethical duty of care to do them no harm.

Sources of data for the insider researcher

Data sources

You will need to decide what constitutes data in your project, and how to gather it.

Sources might include, for example:

Public sources

- Your organisation's public documents, perhaps those published on the organisation's website
- Media texts about your organisation and its work
- Public records (e.g. of parliamentary or judicial proceedings) that discuss your organisation's work
- Governmental policy papers
- Academic studies into the work of your organisation

Sources internal to your organisation

- Widely circulated internal documents, guidelines and manuals
- Limited circulation, 'confidential' internal papers and reports, including board and other committee papers
- Examples of data and paperwork involved in key tasks and activities

Data-gathering activities

- Observations of activities and meetings
- Interviews with your colleagues, perhaps at a variety of levels across the organisation
- Interviews with your organisation's customers and stakeholders (including those critical of the organisation)

Data analysis

- For the insider researcher, the issues are not around how to gain access to data, but how to treat the wealth of data available. You need to work out how to evaluate and weight your data sources, and how these can be represented credibly, while preserving the anonymity of colleagues, and protecting commercial and political sensitivities.
- You may feel that these issues are more problematic for qualitative than for quantitative data, but your organisation may have a wealth of quantitative data that feels like too rich a resource to ignore. When using that data, you will have to judge the extent to which your organisation's data can be taken as representative. You may also find yourself tempted to design your investigations to fit with the available data; this is practical, and entirely understandable, but in evaluating your findings you will have to take care to reflect on the limitations of that approach. In effect, you have to find ways to avoid becoming trapped inside your own data.
- As you collate, analyse and present your data, you need to take particular care to treat each data source appropriately, distinguishing between opinion and evidence-based positions. Your readers will come to their own conclusions about what constitutes solid evidence and what is commercial or political window dressing.
- You will inevitably gather more data than you need. Consider why you choose to report some data to a wider audience and why you choose to keep some for your colleagues or yourself. What are the political implications of the way you focus your story?

Figure B.3: Getting on (1)

Getting on

In the 'getting on' stage, you will be carrying out your research. Textbook after textbook on insider research stresses that it is here that the key strengths

– and weaknesses – of insider research may occur. You undoubtedly know more than an outsider would, but in order to articulate and critically analyse that knowledge, you must, as Coghlan and Brannick advise, 'objectify your subjective experience'. You must find ways to sensitize yourself to your environment, and create a 'strangeness' between yourself and your research subject (Coghlan and Brannick, 2010: 9).

At a practical level, you will need to plan where to begin, what data to gather, and how to gather it.

Questions to consider while you are carrying out your research

You may find it helpful to consider the following questions while you are carrying out your research project. These are challenging questions, but the process of reflection, should, in itself, be helpful.

Objectivity and credibility

How will you create 'strangeness' between yourself and your research topic? How will you help yourself to see the topic with a fresh viewpoint?

- Will this study evaluate your own effectiveness or a method to which you are committed? If so, how will you protect yourself from the temptation to see what you hope to see?
- How will you examine and counteract your own pre-existing biases? Are there creative problem-solving or workshop techniques that you can use? Are there colleagues in other parts of the organisation who can help you with these?

One way to counteract your own biases is to include multiple viewpoints, and ensure that some of your findings come from observers who do not share your assumptions. How will you access multiple perspectives?

- What data will be contributed by others?
- How have you arranged with colleagues or other participants for recognition of their contribution?
- How are you negotiating authorship and ownership?

Power and hierarchies

How will you deal with issues arising from power relationships?

- What steps will you take to avoid coercing (or simply assuming cooperation from) colleagues more junior than yourself?
- How will you ensure that less powerful colleagues don't tell you what they think you want to hear?
- What steps will you take to withstand coercion from colleagues more senior than yourself?
- Remember that power imbalances, direct and indirect, may affect ethical dimensions of your study; you will need to plan how to deal with these.

Are you clear about who needs to give consent for your study?

- Who gives consent in the context of insider research?
- When and how is that consent obtained or assumed?
- Have you explained the implications to all colleagues who will take part in your study, or only the senior colleagues?

How safe do you feel in this institutional environment pursuing this research?

- Is there protection for your interpretations and critical analysis? Can you protect yourself from pressure to report favourably?

Who is responsible for and who is accountable for the final report?

- Will colleagues or committees review your report in draft?
- Are they, and you, both clear about the roles and responsibilities in this regard?
- Who gets final say on what goes in the report?

Who will read the final report or hear the findings?

- Will conflicts arise from your personal relationships with the research subjects?
- Is there potential for conflict to arise from the power relationships in the audience?
- What about an external audience?

Figure B.4: Getting on (2)

During this stage of the project, you may fall into the trap of assuming that because you have gained senior staff or committee approval for the work, you now need only get on and do it. As the BPS notes, 'consent should be an ongoing process and [that] a fuller appreciation of the research and the nature of participation will often become more apparent to participants during the course of their involvement with the research' (BPS, 2014: 21). It

is your job as researcher to keep communicating and negotiating about your research methods and how your findings may be used: securing colleague (including senior colleague) support is not a one-off task, but an ongoing process, which will require a significant investment of your time. It may be tempting to view this as wasted time, or as a progress-blocker, so it is important to remind yourself, too, that investing this time will reap benefits in terms of being more assured that your project will reach completion and be able to achieve impact.

To add value to the field, your research project will have to open up issues for critical enquiry and discussion; this may be perceived as challenging the value system of your organization or professional field in some way. There may be personal and interpersonal challenges. You will need to consider your positioning as a researcher, as an exam board employee and as a colleague, acquaintance or friend.

Getting out

As an insider researcher, you cannot get out of the research context in the same way that a participant observer could if the observer was only temporarily part of the organization under study. Unless, like Bruce Moore, you are willing to resign your position, you are not going to get out physically, so to protect yourself and your colleagues, all the involved parties need to be clear when data gathering is happening and when it is not happening. You need to agree a deadline for the closure of your data-collection processes, and you need to communicate that deadline to all affected colleagues.

You may wish to signal the end of the data gathering, and perhaps the end of your research project, with some sort of event or meeting in which you share your findings with colleagues. As well as marking a clear closure point, this also serves the useful purpose of debriefing the participants and other potentially affected colleagues.

The BPS's *Code of Ethics and Conduct* (2009) includes standards for debriefing research participants, advising that psychologists should:

> (i) Debrief research participants at the conclusion of their participation, in order to inform them of the outcomes and nature of the research, to identify any unforeseen harm, discomfort, or misconceptions, and in order to arrange for assistance as needed.
>
> (ii) Take particular care when discussing outcomes with research participants, as seemingly evaluative statements may carry unintended weight (BPS, 2009: 20).

Once your research is complete, colleagues who have participated in it or are affected by it should have an opportunity to hear about the research and to discuss the findings and conclusions. Staff at all levels of the organization may read evaluative statements as criticisms of their work and find this threatening. Even if you think these are phrased positively, they may imagine implications that involve job loss or changes to working practices that they find alarming. Don't assume that scientific conventions and language will come across as objective, either: to people not used to reading or hearing such language, it probably sounds cold at best and downright harsh at worst. Initiating change may not be the purpose of your project – that does not mean that colleagues will not see it that way and react accordingly. You might need to protect yourself from the potential hostility, but more importantly, you need to protect your colleagues by being very careful about how you express your findings and conclusions.

How to close your project successfully

It is at the close of your project that things are most likely to go wrong, but some simple steps can help to avoid many of these issues. These steps are all essentially about good communication with colleagues.

- Agree a deadline to stop collecting data, and stick to this. Make sure that colleagues know when you are no longer gathering data on the topic. You're not leaving the organisation, so they need to know when you are 'wearing a different hat'.
- Consider a meeting or series of meetings in which you share your findings and conclusions with colleagues, including research participants, senior staff, and anyone else in the organisation with an interest in – or potentially affected by – your research.
- Think about the outputs of your research and how to tailor these to different audiences and purposes. Don't assume busy colleagues will read (or understand and absorb) a research report written using academic conventions.
- Mind your language. You will be using at least two or three levels of specialised language (e.g. the language of research, the technical language of standard setting or assessment, the internal language of your organisation) and for any given audience, even within your own organisation, one or more of those may come across as jargon, or even seem completely nonsensical. Be especially wary of attaching specialised meanings to terms that may be in more general use.
- Be careful about the statements you make, particularly about evaluative language. Remember that you may know that no criticism is intended or change envisaged, but colleagues will not necessarily assume that to be the case, and may be alarmed by your findings and conclusions. You may find it helpful to ask a trusted non-research colleague to review your work and tell you how they think other colleagues will react.
- If you are writing a formal report, or presenting your findings in a formal presentation, don't assume that using scientific conventions and language will render your findings emotionally neutral. In fact, quite the opposite might be true. Readers not used to reading scientific language may not experience it as detached, impartial and fair; instead, they may experience it as judgemental, blunt and cutting.
- If you have feedback that colleagues may experience as negative or critical, it is especially important to think carefully about the form you present it in, the forum for presentation, and the language you use. In order to achieve maximum impact with minimum harm, you may have to think of your findings less as research and more as management feedback. The principles that apply in people management situations to handling negative evaluations constructively apply equally to research findings. Even better, if you have managed the research work collaboratively with affected colleagues, then sharing findings should be less about you giving feedback and more about the project partners discussing the findings.

Figure B.5: Getting out

Getting back

For an exam board practitioner who leads the occasional research project, 'getting back' may seem a simple process of going back to the day job. For the exam board researcher, 'getting back' from any individual project means closing off that project and moving on to another research project. In both cases, the situation is not as simple as it may seem – we should remember Bruce Moore's warning:

> By giving in to the temptation to taste my own guiding assumptions and preferences I had forsaken the luxury of being able to see the world from an epistemologically privileged position. I found that the basis and foundations for my previous understanding and identity had been removed (Moore, 2007: 34).

Researching your own organization, whether in a one-off project or on an ongoing basis, can be a profoundly unsettling experience. You may question your own assumptions, or you may find yourself critical of some of your colleagues' guiding assumptions. Either way, it does not make for a comfortable working environment, and it will not necessarily be helpful when you start to plan your next research project.

To be most effective, insider researchers should consider reconceptualizing their research task. The suggestions captured in these guidelines build on Habermas's theory of communicative action and sophisticated action research methodologies, and emphasize that at all stages of your insider research project, the more time you make for communication and negotiation with colleagues – and the more you see the process as collaboration – the greater your project's chances of success. If all of your interactions in setting up/selling and carrying out the project are cast as collaborative actions, and you reinforce or reiterate this wherever and whenever needed, you will counter any impressions that your project is somehow inspectorial or regulatory, or otherwise sitting in judgement on your colleagues' work. Planning, implementing and communicating about your work in this way will create an impression that your research is conversational and collaborative. While not all colleagues will want, or have time, to be active participants in your research, it may help you as an insider researcher to think of all of your research as participatory, and every piece of research as a joint venture with colleagues.

Acknowledgements

These guidelines were first published as Part 2 of Oxford University Centre for Educational Assessment Report OUCEA/17/2 (2017). The work was supported by a Higher Education Innovation Fund Knowledge Exchange grant, number 1609-VPRC-248.

References

American Friends Service Committee (1955) *Speak Truth to Power: A Quaker search for an alternative to violence*. Online. http://quaker.org/legacy/sttp.html (accessed 19 June 2018).

Bell, J. (2005) *Doing Your Research Project: A guide for first-time researchers in education, health and social science*. 4th ed. Maidenhead: Open University Press.

Blaxter, L., Hughes, C. and Tight, M. (2006) *How to Research*. 3rd ed. Maidenhead: Open University Press.

BPS (British Psychological Society) (2009) *Code of Ethics and Conduct: Guidance published by the Ethics Committee of the British Psychological Society*. Leicester. Online. https://goo.gl/MBo9RQ (accessed 19 June 2018).

BPS (British Psychological Society) (2014) *Code of Human Research Ethics*. 2nd ed. Leicester: British Psychological Society. Online. https://goo.gl/93BCDA (accessed 19 June 2018).

Coghlan, D. and Brannick, T. (2010) *Doing Action Research in Your Own Organization*. 3rd ed. London: SAGE Publications.

Cohen, L., Manion, L. and Morrison, K. (2007) *Research Methods in Education*. 6th ed. London: Routledge.

Costley, C., Elliott, G. and Gibbs, P. (2010) *Doing Work Based Research: Approaches to enquiry for insider-researchers*. London: SAGE Publications.

Denscombe, M. (2010) *The Good Research Guide: For small-scale social research projects*. 4th ed. Maidenhead: Open University Press.

Kemmis, S., McTaggart, R. and Nixon, R. (2014) *The Action Research Planner: Doing critical participatory action research*. Singapore: Springer.

Moore, B. (2007) 'Original sin and insider research'. *Action Research*, 5 (1), 27–39.

Zeni, J. (1998) 'A guide to ethical issues and action research'. *Educational Action Research*, 6 (1), 9–19.

Index

References to tables and exam questions are in *italics*.