EDITORIAL INTRODUCTION

Paul Newton, Jo-Anne Baird, Harvey Goldstein, Helen Patrick and Peter Tymms

England operates a qualifications market in which a limited number of providers are accredited to offer curriculum-embedded examinations in many subject areas at a range of levels. The most significant of these qualifications are:

- the major 16+ school-leaving examination, the General Certificate of Secondary Education (GCSE)
- the principal 18+ university selection examination, the General Certificate of Education, Advanced level (A level).

Although some examinations are offered by only a single examining board, the majority are offered by more than one. And, although each examination syllabus must conform to general qualifications criteria, and generally also to a common core of subject content, the syllabuses may differ between boards in other respects. A crucial question, therefore, is whether it is easier to pass a particular examination with one board rather than another. In fact, this is just one of many similarly challenging questions that need to be addressed.

In recent years, central government has increasingly ensured the regulation of public services and, since 1997, the Qualifications and Curriculum Authority (QCA) has been empowered to exercise regulatory authority over the qualifications market in England. One of the ways in which this has been exercised is through commissioning and conducting investigations into comparability, to identify whether a common standard has been applied across different examinations. For both GCSE and A level, examination standards are expected to be comparable when:

- different versions of the same examination are provided, by a single examining board, within the same year (within-board comparability)
- different versions of the same examination are provided, by different examining boards, within the same year (between-board comparability)
- different versions of the same examination are provided, by a single examining board, from one year to the next (comparability over time)
- different examinations are provided, both within and between boards, across subject areas (between-subject comparability).

The rationale for the book

For many decades now, the comparability of examination standards has been monitored using a variety of techniques. Most of these were originally developed, used and refined by researchers working within examining boards. Over that time, the boards have published a large number of investigations into comparability. Many of these were described in two major reviews, covering the periods 1964–1977 and 1978–1984 respectively, Bardell *et al*. (1978) and Forrest and Shoesmith (1985). These reviews focused exclusively upon studies of between-board comparability: the form of comparability that many considered to be the most important.

Given that nothing similar had been prepared since 1985, despite many studies having been undertaken during the intervening period, it seemed that a third review was long overdue. Indeed, something more than this seemed to be called for. During the decades since 1985, responsibility for commissioning and undertaking comparability monitoring studies had fallen increasingly to the regulator. In addition, the early emphasis upon between-board comparability had broadened to highlight other forms, especially comparability over time. Most significantly, though, developments in comparability theory and practice had begun to raise important questions concerning the range of monitoring techniques available.

Although methods which rely upon the judgement of subject matter experts have been present from the outset of comparability monitoring, they have changed significantly over time; as have methods which rely upon the statistical analysis of examinations data. And, whereas an early enthusiasm for statistical methods (based upon the use of common tests or common examinees) waned during the 1980s and 1990s, it has recently been reinvigorated (with the advent of multilevel modelling).

Given this evolutionary course, the first decade of the third millennium seemed to be an appropriate point at which to take stock of issues such as the following:

- to what extent have the trends in different techniques reflected real methodological progress?
- to what extent are the techniques ultimately based upon the same problematic assumptions?
- are certain techniques to be preferred over others?
- to what extent can each of the methods be improved and is there scope for developing entirely new approaches?

Instead of compiling a descriptive summary of monitoring studies, QCA decided to sponsor a more analytical investigation into the techniques upon which such studies are based; hence, *Techniques for monitoring the comparability of examination standards*.

The process of producing this book

This book was commissioned by QCA on the basis of recommendations from its Standards and Comparability Working Group (a research orientated group with representation from the examining boards and regulators). To appoint, guide and quality-assure the work of chapter authors, QCA appointed an editorial board of five members, together representing a range of perspectives – both academic and professional – and reflecting many decades of expertise in comparability monitoring. Following a competitive tendering process, the editorial board subsequently appointed a panel of authors who represented a similar range of perspectives and wealth of experience.

The main aim of the exercise was to produce a state-of-the-art account of techniques used to monitor the comparability of examination standards in England. Authors were appointed to cover specific techniques, with a number of additional introductory chapters to contextualise the exercise. Each author of methodological chapters was challenged with providing a clear discussion of the technique in question; including its logic, history of use, approach, evolution, technical characteristics, strengths, limitations and potential for improvement.

The quality assurance process was extensive, with considerable exchange between editors and authors. This included a two-day workshop, which provided an opportunity for a wider group of assessment experts to engage with the production of the book. The main activity of the workshop involved sessions in which first drafts of the chapters were reviewed. This was partly to provide additional editorial support, given the wealth of expertise of the participants, but also partly to help the authors and editors identify the main themes and issues that were arising, both chapter-specific and more general. Participants in the workshop are listed in Appendix 1. Our thanks are extended to this group, for providing so many valuable insights.

At the workshop, we offered interested participants the opportunity to write a commentary, should they wish to develop further thoughts on any of the chapters. We envisaged that this would facilitate debate over potentially controversial issues, or provide space for adding particularly salient insights concerning the theory or practice of the technique. The commentaries were quality-assured by the editorial board in a similar manner to the chapters.

Finally, although the authors of each chapter were commissioned to a specific remit, and although influenced by a strong editorial process, they were also encouraged to express their own views. As such, they do not necessarily write with a single voice; nor with a similar perspective on the relative strengths and weaknesses of the techniques discussed; nor even with the same perspective on comparability itself, come to that.

The purpose of this book was at least as much to pose new questions as to answer old ones. And we recognised from the outset that any answers we did provide would not be conclusive. This is not an enterprise characterised by straightforward solutions to problems. Even agreeing the precise meaning of questions can prove problematic, let alone providing rational answers; and providing workable solutions presents yet another level of complexity. We have certainly not reached the end of our journey, but we hope that we have successfully extended the good start made by others. This book helps to tell the story so far.

The audience

As work progressed, it became increasingly clear that we needed to sharpen our remit. Instead of producing a manual which explained everything there is to know about each technique, we decided to produce a handbook that would capture essences: foregrounding underlying conceptual issues and providing references for further reading. By way of illustration, we hope that the book might be useful to:

- enable new assessment researchers to understand the techniques in sufficient depth to be confident in engaging in studies based upon them
- direct any assessment researcher, with responsibility for running comparability monitoring studies, to additional sources which provide full technical details
- provide any educational researcher with sufficient understanding to evaluate studies based upon the techniques.

Although part of the rationale in publishing the book was to promote openness and transparency concerning comparability monitoring, it was written primarily by assessment researchers for assessment researchers. Our aim was to prepare a state-of-the-art review of theory and practice – from a uniquely English perspective – to support practitioners and theorists of the future; both in England and further afield.

The chapters

Following these introductory words the book begins with a general introduction to comparability monitoring in England, written by **Paul Newton**, which sets the scene for the chapters that follow. Newton provides a brief introduction to the qualifications system in England, describing the delivery of the principal school-leaving and university selection examinations. He follows this with a basic explanation of comparability theory. He then highlights the way in which comparability, including comparability monitoring, has increasingly become a central feature of the regulation of qualifications in England.

Kathleen Tattersall subsequently takes us through a historical tour of comparability in England. She traces the roots of England's modern educational and examinations systems back into the 19th century, and to the selection systems of the major universities and the civil service. Tattersall's chapter illustrates how comparability concerns have been a feature of public debate in England for over 150 years. She

highlights three significant stages in the history of comparability in England: the nationalisation of qualifications, i.e. the creation of a single system; the expansion of the system and growth in the number of boards and syllabuses; and, finally, an increasingly formal and powerful regulation of the system.

Colin Robinson introduces comparability from a technical perspective. He focuses on methods for maintaining the comparability of examination standards (rather than methods for monitoring the maintenance of comparability, which are the focus of this book). In addition to explaining present-day practice, he also discusses how we arrived here; highlighting, in particular, debates which have occurred over 'norm-referencing' and 'criterion-referencing'.

The chapter by **Jo-Anne Baird** is the last of the introductory scene-setting ones. Baird teases out what different stakeholders might mean by comparability and considers how these lay views relate to more technical ones. Having identified certain meanings as unsatisfactory – such as those based on pass rates alone – she guides us through a range of more credible definitions: from the technical to the social; from the judgemental to the statistical. Baird finishes by addressing the very thorny question of how to choose between competing definitions.

We then turn to the main substance of the book, with the first of two sets of chapters on techniques for monitoring comparability. These concern techniques that involve applying human judgement to examination papers and performances. **Alastair Pollitt, Ayesha Ahmed and Victoria Crisp** begin this section with a discussion of demands analysis. They explain why the simple question 'is this examination paper more demanding than another?' conceals a very complicated debate on the nature of demand. Fortunately, they do an excellent job of unpicking it, making crucial distinctions between terms such as 'difficulty', 'demands' and 'overall demand'. Having discussed a range of methods for exploring demands, they consider how better to describe, compare and compensate for them.

Robert Adams provides the first of a two-part discussion of what have become known generically as cross-moderation methods. These techniques require subject matter experts, typically senior examiners, to compare examination performances between two or more examinations. Adams focuses particularly upon identification and ratification studies, the most common of the early manifestations. He provides some particularly useful practical insights into how to run this kind of study. He highlights some of the weaknesses of early cross-moderation methods, by way of introduction to the following chapter.

A revolution in cross-moderation methodology occurred with the introduction of the paired comparison technique. **Tom Bramley** introduces this one, tracing its roots to the psychologist Thurstone. The main difference between earlier approaches and paired comparison is that the latter requires judges to make a straightforward overall decision – better or worse – between pairs of scripts from different examinations. These decisions can then be compared to identify a trait of perceived quality, and the average perceived quality of scripts from the different examinations provides some

insight into comparability. This method is especially attractive since it provides a way to control for the severity or lenience of individual judges.

Next, we turn to the second of the two main sets of chapters, exploring techniques which involve the control of background factors and the statistical modelling of results. **Roger Murphy** begins this section with a discussion of methods based upon common tests. Using this methodology, performance on a 'reference' test, exam or element is taken as a proxy measure of the construct in question. Comparing performance in different examinations against performance in the reference test can provide some insight into comparability. Murphy outlines the strengths and weaknesses of techniques based on this principle, but ends on a fairly pessimistic note.

Robert Coe is more optimistic about the use of methods based upon common examinees. Using this technique, common candidate groups are taken to be their own control, such that (on average) the same students might be expected to perform similarly across different examinations. These methods have been used especially to compare standards across different subjects. Despite the substantial technical and theoretical assumptions which need to be made, Coe sees some promise here. He considers these methods to be particularly fit for certain uses of results, particularly when result profiles from various subject areas (e.g. A level physics, mathematics and religious studies) are used to predict performance in distantly related ones (e.g. degree-level psychology).

Taking this section to its logical conclusion, **Ian Schagen and Dougal Hutchison** discuss the use of multilevel modelling. This technique takes statistical control as far as practical measurement will allow. It provides some insight into comparability by investigating the mean grade differences between examinations that remain once a range of significant background variables have been controlled for. Multilevel modelling is a relatively new weapon in the comparability monitoring armoury, and the authors are very positive about its potential.

Finally, the book ends with a conclusion. We consider what we have learned from over half a century of comparability monitoring work, as well as what we are still unsure of. We compare the relative strengths and weaknesses of the various methods, and provide recommendations for future research and practice.

References

Bardell, G.S., Forrest, G.M., & Shoesmith, D.J. (1978). *Comparability in GCE: A review of the boards' studies, 1964-1977*. Manchester: Joint Matriculation Board on behalf of the GCE Examining Boards.

Forrest, G.M., & Shoesmith, D.J. (1985). *A second review of GCE comparability studies*. Manchester: Joint Matriculation Board on behalf of the GCE Examining Boards.

Appendix 1 List of workshop participants

Martin	Adams	Research Machines
Robert	Adams	Assessment and Qualifications Alliance
Angus	Alton	Qualifications and Curriculum Authority
Jo-Anne	Baird	University of Bristol (AQA at the time of the workshop)
Andrew	Boyle	Qualifications and Curriculum Authority
Tom	Bramley	Cambridge Assessment
Robert	Coe	University of Durham
Simon	Eason	Assessment and Qualifications Alliance
Gill	Elliott	Cambridge Assessment
Mike	Forster	Oxford Cambridge and RSA Examinations
Dee	Fowles	Assessment and Qualifications Alliance
Amy	Glassborow	Qualifications and Curriculum Authority
Harvey	Goldstein	University of Bristol
Jeffrey	Goodwin	Edexcel
Elizabeth	Gray	Oxford Cambridge and RSA Examinations
John	Gray	University of Cambridge
Jackie	Greatorex	Cambridge Assessment
Malcolm	Hayes	Edexcel
Sandra	Johnson	Assessment Europe
Mike	Kingdon	Assessment Consultant
Iasonas	Lamprianou	University of Manchester
Alison	Matthews	Qualifications and Curriculum Authority
Michelle	Meadows	Assessment and Qualifications Alliance
Roger	Murphy	University of Nottingham
Paul	Newton	Qualifications and Curriculum Authority
Bruce	Nicholson	Publications Consultant
Isabel	Nisbet	Qualifications and Curriculum Authority
Tim	Oates	Cambridge Assessment
Dennis	Opposs	Qualifications and Curriculum Authority
Helen	Patrick	Assessment Consultant
Liz	Phillips	Welsh Joint Education Committee
Anne	Pinot de Moira	Assessment and Qualifications Alliance
Alastair	Pollitt	Assessment Consultant
Mick	Quinlan	National Assessment Agency
Jonathan	Robbins	The Talent Centre Ltd
Colin	Robinson	Assessment Consultant (QCA at the time of the workshop)
Ian	Schagen	National Foundation for Educational Research
Gordon	Stobart	Institute of Education, University of London
Steve	Strand	University of Warwick
Neil	Stringer	Assessment and Qualifications Alliance
Kathleen	Tattersall	Institute of Educational Assessors
Raymond	Tongue	Welsh Joint Education Committee
Peter	Tymms	University of Durham

Rob	van Krieken	Scottish Qualifications Authority
Colin	Watson	National Assessment Agency
Chris	Wheadon	Assessment and Qualifications Alliance
Chris	Whetton	National Foundation for Educational Research
Alison	Wood	National Assessment Agency (QCA at the time of the workshop)