



Qualifications and
Curriculum Authority

What has England learned from decades of monitoring the comparability of examination standards?

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

Version 1.1, 19 October 2007

Paper presented at the 8th Annual Conference of the Association for Educational Assessment – Europe, 8-10 November 2007, Stockholm, Sweden.

Lead author

Dr Paul E. Newton
Head of Assessment Research
Regulation and Standards Division
Qualifications and Curriculum Authority
83 Piccadilly
London
W1J 8QA
UK

+44(0)2075095601

newtonp@qca.org.uk

Co-authors

Dr Jo-Anne Baird
University of Bristol

Professor Harvey Goldstein
University of Bristol

Professor Peter Tymms
University of Durham

Abstract

England has a very long tradition of monitoring the comparability of examination standards for its major school-leaving and university-entry examinations (comparability between boards, over time, between subjects, etc.). Across a period of over 50 years, a variety of techniques have been developed specifically for this purpose: some of which have prioritised human judgement of scripts; and some of which have prioritised statistical analysis of results.

This paper reports on a major review of techniques for monitoring the comparability of examination standards in England. The review has resulted in the production of an edited book, with chapters devoted to describing and evaluating each of the major techniques, and chapters explaining the context in which comparability has been debated and investigated.

This paper will explain the kind of comparability challenges that have been faced in England, and will describe some of the techniques that have been used to investigate it. It will discuss some of the problems that have been encountered over the years and some of the solutions that have been developed. It will explain some of the tensions and challenges that still remain and will consider the extent to which genuine progress can be said to have been made.

Introduction

Concerns over comparability have been a feature of public debate in England for over 150 years now, since the earliest days of school-based examining. Over 50 years ago, partly in response to concerns such as these, the examining boards began to develop techniques for monitoring comparability, to investigate whether such concerns had any basis in fact.

In the early days of comparability monitoring, it constituted an excellent example of self-regulation: being initiated and undertaken exclusively through collaboration between the examining boards. Nowadays, the qualifications market in England is formally regulated by an independent body – the Qualifications and Curriculum Authority (QCA) – and this organisation has taken on some of the responsibility for monitoring comparability, albeit still in collaboration with the examining boards (who still undertake their own monitoring), and alongside other independent researchers.

Unfortunately, despite 50 years of research in the area, the methods that we use – and even the underlying theory of what we are trying to do – still remains quite controversial. This is why the QCA recently commissioned a state-of-the-art review of techniques for monitoring the comparability of examination standards. It wanted to know:

1. the extent to which our comparability monitoring research is based upon a solid foundation;
2. whether the more complicated techniques that we tend to use nowadays are better than the less complicated ones that we used to use; and
3. whether we should be using certain techniques in preference to others.

The following account provides a brief overview of comparability in England today, and then identifies some of the lessons that were learned from the review (which will be published as the edited volume, Newton, *et al.* 2007).

Comparability monitoring

Comparability is defined here as the application of the same standard across different examinations. It is related to the idea of:

- equating, where the intention is to calibrate tests built to the same content and statistical frameworks; and
- linking, where the intention is to calibrate tests built to different frameworks.

In fact, though, techniques for monitoring comparability are neither techniques for linking, nor techniques for equating, because their purpose is to investigate the defensibility of pre-existing calibrations. They are used to check that comparability *actually* exists where it is *supposed* to exist.

Examinations in England

It is useful to situate this discussion of comparability within the context of England's largest public examination, the GCSE. So here are a few important background details:

- the GCSE is England's principal school-leaving examination, which replaced the O level in 1988;
- students tend to study 8 to 10 GCSE subjects over a period of 2 years;
- some subjects have to be studied – like English, maths and science – but others are optional;
- there is a large range of subjects from which to choose, although not all schools will teach all subjects;
- almost everyone takes at least one GCSE;
- over 5 million examinations are sat each year.

There are three GCSE examining boards based in England (AQA, Edexcel and OCR) and there is one each based in Northern Ireland and Wales (CCEA and WJEC, respectively). However, all of the boards offer similar examinations, and they are all in competition with each other for their share of the 'three-country' qualifications market. This element of competition is important when it comes to understanding why we, in England, feel a need to monitor comparability that may not be felt in other countries.

Comparability in England

The most basic requirement for comparability operates at the level of an individual examination – within a single examining board – as a different version is sat from one year to the next. Most GCSEs, though, have multiple tiers of entry – harder and easier versions of each examination, with some overlap of available grades – so even the basic requirement for comparability is quite complicated. Furthermore, within any particular examining board, there may be more than one syllabus on offer for a particular subject, so the standards in these alternative syllabuses need also to be equivalent.

Clearly, what is true within a single board, also needs to be true across boards. So, where the same subject is offered by multiple boards, the standards for equivalent grades again need to be comparable.

Finally, there are the more complex comparability requirements: between subjects, and over decades in time. In effect, if these demands are taken to their logical conclusion, the ultimate extended implication of comparability is that: a grade C from a 1997 AQA GCSE maths exam (syllabus 1 higher tier) ought to be of the same standard as a grade C from a 2007 OCR English exam (syllabus 2 foundation tier). Exactly what that might mean, though, is far from straightforward to understand.

Methods for linking standards

A few words need to be said on how standards are linked in the first place, which is best explained by beginning with how they are not linked.

Unlike in some countries, examination standards are not linked by cohort referencing (for example, by awarding a grade A to the top 10% of each cohort, a grade B to the next 15%, and so on). Cohort referencing is not at all palatable when substantially different kinds of student enter for substantially different kinds of examination; which is commonplace in England.

Unlike other countries, examination standards are not linked through pre-test linking studies (which use a combination of experimental design and statistical analysis to engineer comparability). On the one hand, there are many practical challenges which are effectively insurmountable in the GCSE context (e.g., too many examinations to link, the high security risk, the fact that examinations are in the public domain once sat). On the other, even if it were possible to overcome these challenges, the rate at which syllabuses and examinations change would often render outcomes from this kind of study quite contestable anyhow.

In England, standards are linked, after the event, primarily through the professional judgement of experienced examiners. In short, once an examination has been sat, and there is solid evidence of how students have performed in it, groups of senior examiners are assembled to review this evidence, and to decide upon grade boundary cut-scores that would link this year's mark scales to last year's. This process is therefore grounded in examiners' perceptions of the relative quality of scripts, at different marks, from one year to the next; their judgements being supported by a range of statistical evidence (which helps them to appreciate how the cohorts under comparison may have differed).

The need to monitor comparability

The preceding context is important to enable a full appreciation of why the English feel the need to monitor comparability at all. The examining boards are not simply trusted always to have done their job properly, for a variety of reasons.

First, there is a recognition that the approach to maintaining standards in England is quite fragile: being fairly subjective (based upon human judgement); and there not being a great deal of coordination between examining boards, even when setting standards for parallel examinations (there is certainly more nowadays, but still not a great deal).

Second, there is a subtle, but widespread, sense of distrust of the qualifications market. If, for example, an examining board happened to offer an easy route to a GCSE then this might well distort the market; and occasionally the boards are explicitly accused of lowering standards to improve their market share. The idea of there being a market for qualifications in England is the critical factor here. In fact, this kind of distrust is not limited to the qualifications market; the qualifications market is just one of many in England which is formally policed by an independent regulator. So there exists a general sense of public distrust, as well as a specific one related to qualifications.

Third, examination standards are repeatedly – year after year, decade after decade – the subject of both lay and academic criticism. Sometimes this is based upon apparently persuasive evidence; and sometimes it is based upon no evidence whatsoever. Yet, the impact can be very destabilising, either way.

All of these reasons help to explain why so much effort goes into monitoring comparability. Importantly, the purpose of monitoring is primarily formative, in the sense that any evidence of discrepancy is used to rectify standards during the following examination session.

How is comparability monitored?

There are essentially two approaches to monitoring comparability: one largely judgemental and one largely statistical. Judgemental methods have always been the mainstay of comparability monitoring, while statistical methods have tended to be more controversial, and have fallen in and out of favour over the years. From within both of these perspectives, though, there have been trends over time in preference for different techniques.

From the judgemental perspective, much of the history of comparability monitoring has been dominated by the ratification method. During the late 1990s, though, this method was superseded by the paired comparison. Similarly, from the statistical perspective, although there was a flurry of interest in the use of reference test methods from the late 1960s to the late 1970s, their use had been largely discontinued by the mid 1980s. Since then, they have only been used infrequently. From the mid 1990s onwards, though, the boards began to take interest in a new statistical approach to monitoring comparability based upon multilevel modelling. In short, there have been some very clear, and discrete trends in the history of techniques for monitoring comparability in England. This raised an important question for the review: to what extent did these trends reflect genuine technological progress? Or, more straightforwardly, are the new techniques really any better than the old ones? This will be considered through discussion of the current favourite judgmental and statistical methods.

Paired comparison

A typical investigation based upon the paired comparison method would explore the comparability of standards across versions of an examination offered by different boards (e.g., different GCSE mathematics examinations). The essential steps in the exercise might be as follows:

1. select one mathematics syllabus from each board (probably the largest-entry one);
2. select five scripts from each syllabus, at each of two grade boundaries (only two grade boundaries tend to be chosen, normally A and C, due to the large number of judgements required for such a study);
3. select three senior examiners from each board to do the judging (ideally, the most senior ones);
4. devise a plan for the examiners to judge pairs of scripts from different boards (so that enough examiners judge enough of the pairs, and to ensure that they are not asked to judge scripts from their own board);
5. for each pair of scripts under comparison, ask the examiner to judge which is of a higher quality (requiring them to do so quickly, without allowing ties);
6. once all script-pairs have been judged, analyse the results of those judgements using Rasch software.

This process will result in a single scale – for each grade boundary under review – that represents the perceived quality of each script, relatively speaking. Recall that *all* of the scripts on the scale will have come from the same grade boundary – so they are all *meant* to be of the same standard – but the scale will represent any perceived differences between

them. If these estimates are then averaged within boards, the differences between board averages will imply differences in the perceived quality of their borderline scripts, i.e., perceived differences in grading standards.

The paired comparison method has a couple of major advantages over other techniques that have been used in the past. In particular, the examiners are not required to hold an absolute standard in their head, because they are only asked to make judgements of relative worth. In addition, though, as a consequence of the Rasch analysis, this automatically provides information on mis-fitting scripts and judges, which can help the researcher to evaluate the validity of the exercise. In short, although the methodology can end up being quite time-consuming (and a bit boring) for judges, it does appear to represent a substantial technological advance over earlier judgemental techniques.

However, as with all of the judgemental methods that have been used over the years, it is still somewhat open to question. Unfortunately, it is a fact of educational assessment that quality of task performance is highly context-sensitive: questions that look very similar to senior examiners may prove to be very different in difficulty for students. And we know from research that even experienced teacher-examiners are often quite bad at judging question difficulty; let alone at adjusting their perceptions of task performance to control for differences in task difficulty. Yet, the basic assumption of any judgemental method is that examiners will be able to do exactly this, with reasonable precision.

Again, the Rasch analysis does routinely provide information on consistency of judgements, which can be reassuring when seen. However, consistent judgements are not necessarily accurate judgements. And judgemental methods can be quite susceptible to systematic bias. In short, examiners may well be able to provide researchers with decisions – even consistent decisions – but do those decisions necessarily say much about comparability?

Multilevel modelling

Multilevel modelling is a generalisation of multiple regression, which can actually be seen as an extension of the use of reference tests to monitor comparability.

During the 1970s, it became very evident that reference-test-based comparability monitoring studies suffered from a major limitation: the reference test inevitably failed to measure all of the factors that led to successful performance in the examinations under comparison. When reference tests were used, stakeholders would often respond along the lines of:

- no account was taken of student motivation, and this syllabus is much more motivating than that one; or
- no account was taken of teaching quality, and teachers of this syllabus are much better than teachers of that one.

And they were right to voice these kinds of concern; which rendered results from reference test analyses entirely indeterminate, i.e., not useful.

More recently, researchers have turned to multiple regression techniques, and to multilevel modelling in particular. The 'holy grail', here, is to measure (either directly or by proxy) all of the variables that affect attainment. If all of the 'input' variables are measured adequately, then it should be possible to predict the 'outcome' measure, the examination result, with confidence. Having done so, differences between boards, between predicted and actual results, would indicate differences in grading standards. The logic is very attractive. It almost seems to offer the potential for the ultimate comparability monitoring study.

From the perspective of statistical control, the use of multilevel modelling clearly represents a major technological advance over the use of simpler methods, such as reference tests. Of course, to represent a major practical advance, it would be essential to measure those factors that the simpler methods do not. This has not always been the case, especially when the multilevel modelling has focused principally upon the analysis of prior or concurrent attainment results (ignoring factors like motivation, effort, teaching quality, and so on). More importantly, still, as long as there are *any* key factors left uncontrolled – teaching quality, for example – the analysis is still legitimately open to challenge. Again, however statistically sophisticated the analysis, and however many of the key factors are accurately measured, if *any* of the key factors remain uncontrolled, then the conclusions from the analysis will still be somewhat indeterminate.

Judgemental or statistical methods?

The recent zeitgeist, of once again finding statistical methods attractive for monitoring comparability, raises an important question: ought we to put more confidence in results from judgemental or statistical methods? Unfortunately, the review did not reach a conclusion on this matter, since both had their strengths and weaknesses.

In a similar situation, two decades ago, a previous review had come down strongly in favour of judgemental methods. Its justification was that these are most close to the methods used to set standards in the first place, during awarding meetings. However, although this was –

and still is – undoubtedly true, precisely the opposite argument could equally be made: when monitoring comparability, we might actually be better off using non-judgemental methods, to avoid being led astray by exactly the same judgemental biases that have the potential to compromise awarding meetings. At the very least, the decision is not obvious, which probably recommends using both types of approach, wherever possible.

Conclusion

The review resulted in a number of conclusions:

1. we *have* made genuine technological progress in the development of methods for monitoring comparability (both in terms of judgemental methods and in terms of statistical methods);
2. these methods *do* have the potential to provide at least reasonably defensible insights into comparability (even if our conclusions are necessarily tentative and open to debate);
3. we are less clear now, than in previous decades, over whether to prioritise judgemental or statistical methods (so it probably makes sense to use both, wherever possible).

However, perhaps the most important conclusion from the review was that our biggest challenge is not technological, but conceptual. After 50 years of research into comparability monitoring in England, the most significant unresolved issue is not whether one technique, or class of techniques, is better than any other. It is how best to specify, and to accommodate, alternative conceptions of comparability. There are at least two major facets to this:

1. what do the professionals – the examining boards and the regulator – think that they mean when they pronounce that standards are comparable?
2. perhaps more importantly, what do users of examination results think that they are doing when they draw inferences from examination results for different purposes?

The crucial issue, stemming from this second point, is that different uses of results implicate different conceptions of comparability. Not all uses necessarily implicate a specific conception. However, certain uses certainly do; and, moreover, certain of the conceptions implicated by these uses are mutually incompatible. This raises an interesting question for debate. How ought we to monitor comparability when:

- results are used for many different purposes – from university selection, to system monitoring, to personal qualification – certain of which invoke mutually incompatible conceptions of comparability;
- there is no official, nor *de facto*, prioritisation of purposes;
- different stakeholders subscribe to different conceptions of comparability (but often fail to appreciate this);
- there is no official statement of which conception ought to be operationalised when setting standards (i.e., during grade awarding);
- even the academics are unable to agree upon a system for classifying conceptions of comparability, or to agree which conceptions might be legitimate, or illegitimate, from a theoretical perspective;
- and so on?

These are challenges that will need to be faced during the next 50 years of comparability monitoring research.

References

Newton, P.E., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (2007) (Eds.). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

[For a more detailed discussion of the issues raised in this presentation, including a full list of supporting references, see the concluding chapter of the edited volume above.]