Techniques for monitoring the comparability of examination standards

Paul Newton
Qualifications and Curriculum Authority

Version 2.1, 16 August 2007

Paper presented at the International Association for Educational Assessment 33rd Annual Conference, 16-21 September 2007, Baku, Azerbaijan.

Author details

Paul Newton is Head of Assessment Research in the Regulation and Standards division of England's regulator for national qualifications, where his work focuses on issues related to the design and evaluation of large-scale educational assessment systems.

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

Abstract

England operates a qualifications market in which a small number of examining boards are accredited to offer relatively 'authentic' curriculum-embedded examinations in a range of subject areas at a range of levels. Within this context, questions naturally arise as to whether examination standards are comparable: from board to board; from year to year; from subject to subject; and so on.

This paper describes the outcomes of a major review into techniques for monitoring the comparability of examination standards; techniques which have been employed by the examining boards and regulatory authorities, in England, over the past fifty years or so. It explains how a range of different techniques have been developed and identifies an interesting evolutionary history. Trends in preference for judgemental and statistical approaches are considered in terms of systemic, social, technical and conceptual factors. Consideration is given to the extent to which progress has been made; and some of the challenges which still remain are highlighted.

Introduction

This paper concerns comparability – the application of the same standard across different examinations – and techniques for monitoring it. It is becoming increasingly common in educational measurement texts to distinguish between:

- 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, this paper is neither about linking, nor about equating, because the techniques in question are used to investigate the defensibility of pre-existing calibrations. These are methods used to *monitor* comparability, not to *create* comparability. They are used to check that comparability *actually* exists where it is *supposed* to exist. Having said that, the kind of comparability we are talking about is often, but not always, more on a par with linking than equating.

In England, formal monitoring exercises have been conducted for over 50 years now, generally focusing upon our major school-leaving and university-selection examinations, and using various different investigative approaches. However, despite 50 years of research, the methods that we use, and even the principles underlying what we are trying to do, remain quite controversial.

For this reason, the Qualifications and Curriculum Authority – England's national qualifications regulator – commissioned a state-of-the-art review of techniques for monitoring the comparability of examination standards. We were keen to know:

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

This paper presents a summary of the outcomes of that review (which will be published as Newton, *et al.* 2007). It begins by introducing England's largest examination, then explains the ramifications of comparability in England, and then describes the techniques and their evolutionary history.

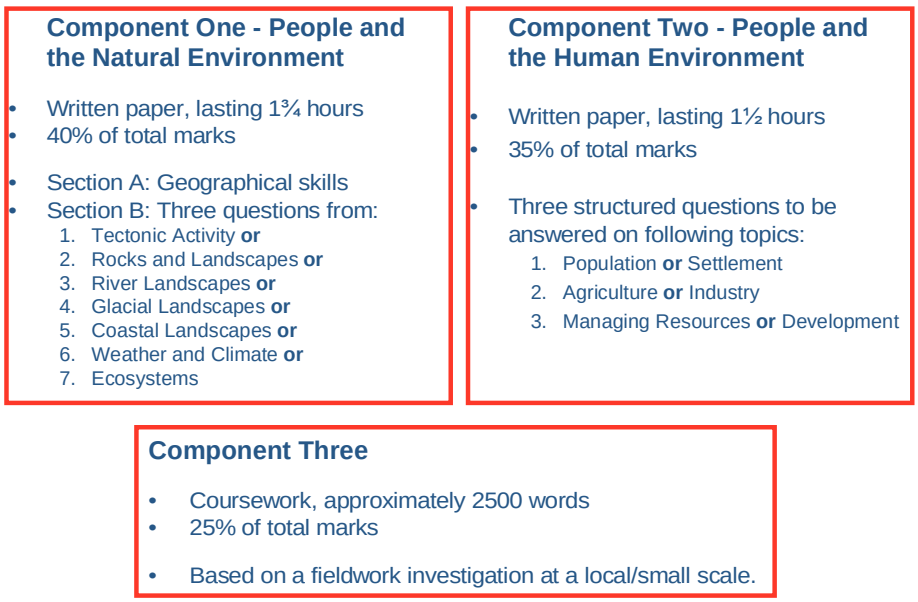
The General Certificate of Secondary Education

The GCSE is England's major school-leaving examination. It replaced the General Certificate of Education O level in 1988, offering an examination aimed at all school-leavers, not just the highest-attaining ones. Students tend to study 8 to 10 GCSE subjects over a period of 2 years. Some of these are compulsory – like English, mathematics and science – but others are optional. Almost everyone takes at least one GCSE examination, and over 5 million are sat each year.

There are three GCSE examining boards based in England: AQA, Edexcel and OCR. Northern Ireland and Wales have one each: CCEA and WJEC, respectively. However, all of the boards offer similar examinations, and are in competition for their share of the qualifications market across the three countries. So, for example, quite a lot of students in Wales and Northern Ireland sit examinations from one of the English boards. This issue of competition between boards is important, with implications for perceptions of comparability.

GCSEs are offered in a very wide range of subjects; from astronomy, to drama, to manufacturing, to Welsh as a second language. They are all fairly similar in examination structure, though, and Figure 1 illustrates a typical example, which happens to be a geography syllabus from AQA.

Figure 1 A geography syllabus from AQA



In common with many GCSE examinations, this example from AQA geography comprises three components: two written papers and a coursework element. The written papers include mainly constructed-response questions, either short- or long-answer, and the coursework is written up as a project of around 2,500 words.

What is not apparent from Figure 1 is that most GCSEs work on a principle of differentiated assessment. This means that higher- and lower-attaining students have alternative written papers, to ensure that all students are appropriately stretched. There are normally two tiers of entry:

1. the higher tier is targeted at students who are likely to achieve grades A* to C/D; and
2. the lower tier is targeted at students who are likely to achieve grades C/D to G.

Comparability

Mathematics has traditionally been one of the few examinations – latterly the only examination – with three tiers of entry. In fact, the new syllabuses have only got two, but it makes the point well to introduce the principle of comparability using the old three-tier structure. In the context of GCSE mathematics – even within single syllabus of a single board from one year to the next – comparability requirements are very demanding. For example, at the grade C boundary:

- the year 1 higher tier has to link to the year 2 higher tier; but also
- year 2 higher tier has to link to the year 2 intermediate tier.

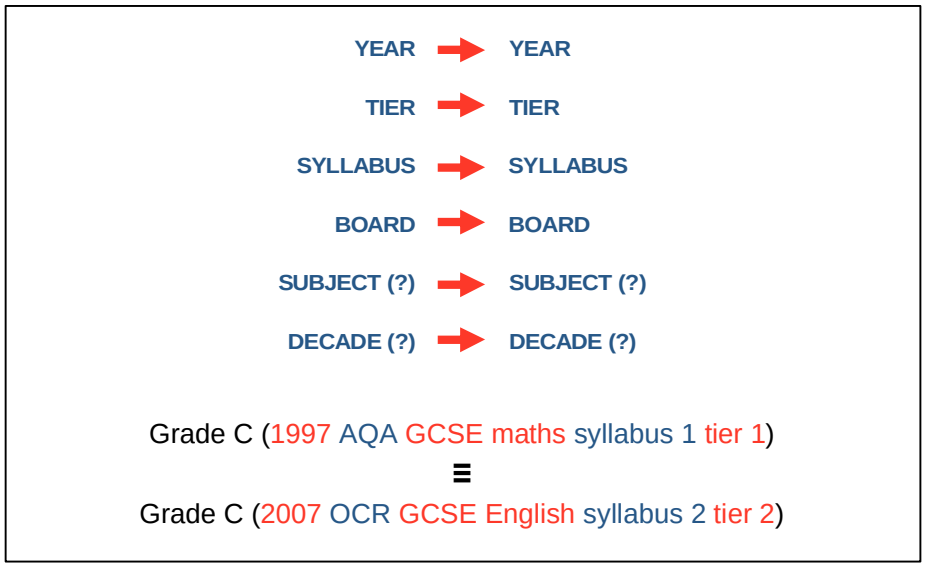
Similarly, at the grade D boundary:

- the year 1 intermediate tier has to link to the year 2 intermediate tier; but also
- the year 2 intermediate tier has to link to the year 2 lower tier.

And so on, for all the other grades, with multiple links needing to be created simultaneously. But that is not the end of the story, because most of the boards have more than one mathematics syllabus. So, within each of the examining boards, all of their mathematics syllabuses need to apply exactly the same standard, for each grade boundary. Moreover, all of the mathematics syllabuses, within all of the five examining boards, need to apply exactly the same standard, for each grade boundary. During 2006 (ignoring pilot syllabuses and the like) there were 12 GCSE mathematics syllabuses; all of which had to apply exactly the same standard, for each grade boundary. Of course, this is just within a single subject area,

from one year to the next. In reality, things are more complicated still, as indicated in Figure 2.

Figure 2 An illustration of the extended implications of comparability



In the system that we operate in England, we require that standards (at equivalent grades) be linked: from year to year; from tier to tier; from syllabus to syllabus; and from board to board. For the first three of these requirements, we have a code of practice which specifies exactly how standards ought to be linked. To help with the fourth requirement, the boards – through their Joint Council for Qualifications – have become very proficient at sharing background data on cohort characteristics. Beyond the fourth requirement, the extended implications of comparability are more vague. However, there is a general expectation that standards ought to be comparable, both between subjects and over extended periods of time. In a sense, they have to be, to make sense of aggregated results within school performance tables: comparability between subjects is what allows us to compare schools whose students have sat examinations in different subjects; comparability over extended periods of time is what allows us to monitor trends in school effectiveness. So, in a sense, the ultimate extended implication of comparability is that a grade C from a 1997 AQA GCSE mathematics examination (syllabus 1 higher tier) is of the same standard as a grade C from a 2007 OCR English examination (syllabus 2 foundation tier).

Creating comparability (linking standards)

Although this paper concerns methods for monitoring comparability, not methods for creating comparability, it is important to say a few words on how standards are linked in the first place. In England, standards are linked 'after the event' by professional judgement. That is, once an examination has been sat, and solid evidence has been collected concerning how students have performed in that examination, we assemble a group of senior examiners to decide grade boundary cut-scores that would link this year's mark scale to last year's. This happens for each syllabus separately. As such, the process is based upon examiners' perceptions of the relative quality of scripts, at different marks, from one year to the next; and it is supported by a range of statistical evidence, which helps the examiners to appreciate how the cohorts under comparison may have differed.

The need to monitor comparability

This background information is important for explaining why England feels the need to monitor comparability at all. As a society, we do not solely trust the examining boards to have done their job, as far as comparability is concerned. This is true for numerous reasons.

First, it is widely recognised that the approach adopted in England for creating comparability is quite fragile:

- it is fairly subjective, being based upon human judgement;
- there is not a great deal of coordination between examining boards, even when setting standards for parallel examinations (admittedly, there is more coordination nowadays, but there is still not a great deal);
- more problematically, there is no consensus over how to achieve certain forms of comparability (particularly between examinations in different subject areas).

In a sense, then, there is an acceptance that the process of grade awarding is open to review, very much as legal decisions are.

Second, there is a subtle, but widespread, sense of unease with the qualifications market. If, for example, a board happened to offer an easy route to a GCSE then this might well distort the market. Indeed, on occasion, the boards are explicitly accused of lowering standards to improve their market share. In fact, this kind of unease 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 unease, as well as a specific one related to qualifications.

Third, examination standards are repeatedly – year after year, decade after decade – the subject of criticism from both lay and academic stakeholders. 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 we, in England, put so much effort 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.

Techniques for monitoring comparability

Given the need to monitor comparability, how is this undertaken? The following sections will describe four different approaches to monitoring comparability, two largely judgemental and two largely statistical. Judgemental methods have always been the mainstay of comparability monitoring, while statistical methods have tended to be more controversial. From within both of these perspectives, though, there have been trends over time in preference for different techniques. Thus, from a judgemental perspective, much of the history of comparability monitoring has been dominated by the ratification method. Yet, during the late 1990s, this was completely superseded by the paired comparison method. Similarly, from a statistical perspective, although there was a flurry of interest in common test methods from the late 1960s to the late 1970s, their use had been largely discontinued by the mid 1980s. Since then, they have been used only infrequently. Yet, towards the mid 1990s, researchers began to take an 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. An important question for our review, therefore, was whether or not these trends reflected genuine technological progress; that is, whether the new techniques were really any better than the old ones.

The following sections will describe each method in the context of a between-board comparability study, in which a single subject (e.g., GCSE geography) is the focus of attention, and for which the highest-entry syllabus (for each board) is under scrutiny.

Ratification method

The logic of a between-board ratification study is quite straightforward:

- assemble a group of senior examiners in the same room (around 3 from each board);
- ask them to scrutinise grade boundary scripts from the different examinations (around 4 each, from a sample of 10);
- for each grade boundary script, ask them either to 'ratify', or to 'repudiate'.

In ratifying a script, an examiner is agreeing that it represents work of an appropriate standard. In repudiating a script, an examiner is claiming that it represents work of a higher or lower standard.

The examiners do not come to this task entirely unprepared, since they will have been provided with a full complement of relevant syllabuses, papers and mark schemes in advance. They are asked to judge the quality of performance that they see in each script in the context of its relevant syllabus, paper and mark scheme. Examiners are not asked to judge the quality of scripts from their own board. Instead, they are generally asked to assume that the standard in their own head – presumably the standard that was applied to their own board's examination – is the standard against which the other boards' examinations ought to be judged.

By way of summary, for a typical ratification study:

- 1 syllabus selected from each board
- 10 scripts available from each syllabus, at each grade boundary
- scripts selected from the A and the C boundary, respectively[1]
- each script represents the complete work of a student at the boundary mark
- exercise repeated separately for each boundary:
    - 3 examiners, per board, judge quality of scripts from *other* boards
    - each script judged as: on (0), below (-), or above (+) 'the standard'
    - each examiner judges around 4 scripts per syllabus, per boundary
    - overall, with 15 examiners, 48 judgements per syllabus, per boundary (240 judgements in all)

[1] It is not feasible to repeat the task for all boundaries, so only those deemed most important are studied.

The analysis of these data is fairly crude; little more than a comparison of frequencies. But the patterns of results can still be illuminative, particularly when all examiners are in agreement.

The main strength of this method is that it requires the most experienced examiners from the country – those who are actually empowered to set standards in their respective boards – to decide whether they are all singing from the same hymn-sheet, as far as standards are concerned. The main weakness of this method is that it requires them to "to spot a borderline script at twenty paces" – as Tom Christie and Gerry Forrest once commented – which might not be that easy in the context of an unfamiliar syllabus, paper and mark scheme. Another, more practical, problem is that it is hard to quantify exactly how far out of line any discrepant board is. It is therefore equally hard to decide what action ought to be taken to re-align standards.

Paired comparison method

The paired comparison method was designed specifically to overcome certain limitations of the ratification method. It is essentially the same, although, instead of requiring examiners to judge whether a single script is of the appropriate standard, it simply requires them to judge which of two scripts (from two examinations) is of a higher quality. This requires them to make a judgement of relative worth, rather than of absolute grade-worthiness. Despite the fact that all scripts are supposed to be of exactly the same standard, the judges are asked to give a 'gut reaction' as to which strikes them as the better (and they are forced to choose one, since there are no ties allowed).

Again, by way of summary, for a typical paired comparison study:

- 1 syllabus selected from each board
- 5 scripts available from each syllabus, at each grade boundary
- scripts selected from the A and the C boundary, respectively
- each script represents the complete work of a student at the boundary mark
- exercise repeated separately for each boundary:
    - 3 examiners, per board, judge quality of scripts from *other* boards
    - scripts judged in pairs – one board against another – to identify the 'higher quality' script in each pair (no ties allowed)
    - each examiner takes around 4 minutes per pair, judging around 75 pairs
    - overall, with 15 examiners, around 1,125 judgements per boundary

- Thurstone analysis (using Rasch software) estimates 'judged difficulty' of each script

The analysis of these data is far more sophisticated, since the method is amenable to the use of Rasch. This produces an estimate of judged difficulty for each script. When these estimates are averaged, within boards, the overall pattern can identify whether certain boards are more lenient or harsh than others.

The main strength of this method is that examiners do not have to internalise grade boundary standards. But it also has the advantage of providing statistics on mis-fitting scripts and judges, which can help to validate the process. The main weakness of this particular method is that the task requires so many judgements that it can become very tedious and tiring for examiners.

There are also more general limitations of judgmental methods, which need to be recognised. Technically speaking, a major drawback is the potential un-representativeness of the process, with so few scripts and so few examiners involved. But a more serious issue is whether the task is simply too complicated for examiners to perform accurately (particularly when faced with examinations built to different content and statistical frameworks – exams they are not familiar with – and particularly when they are required to make their decisions so quickly). The examiners may provide us with results – even consistent results – but do those results necessarily tell us much about comparability?

Common test method

The logic of the common test method is also straightforward, albeit from a statistical perspective. In short, it says: given students of a similar calibre, we should expect similar examination results. To estimate calibre, we use a reference test (sometimes a common component, sometimes an entirely distinct test). And to express the relationship between calibre and examination results, separate regression lines are calculated for each board. If, across boards, a similar relationship is observed – between student calibre and examination results – then the boards are assumed to have comparable standards.

Essentially, common test methods are based upon a principle of statistical control. So, for example, the use of an aptitude test would control for the impact of aptitude upon attainment, for the respective examination cohorts. Unfortunately, this leaves this method open to the challenge that examination cohorts may differ in terms of other variables, which genuinely

impact upon attainment but which are not controlled for. For example, one cohort might have been taught better than another, or one cohort might have studied for longer than another.

So, on the one hand, this method can be extremely economical in an examinations context; particularly when the common test is a common component. Yet, on the other, it is so easy to mount a plausible challenge – on the basis of uncontrolled variables – that it can be hard to have any confidence in the results, at all. This is basically why examining board researchers largely stopped using these methods to monitor comparability in the 1980s.

Multilevel modelling methods

Of course, statistical methods are not necessarily restricted to modelling the impact of single variables. For some time now, regression methods have allowed us to investigate the combined impacts of multiple variables. Recently, with the introduction of multilevel models – which are able to accommodate the kind of cohort clustering which is common with examinations data – there has been a flurry of interest in using multiple regression techniques for monitoring comparability.

The holy grail, here, would be 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 here. It seems to offer the potential for the ultimate comparability monitoring study. However, you do need to be able to measure the previously uncontrolled variables, before you can include them into your multilevel model. And that can be very problematic. More importantly, as long as *any* of the key variables remain uncontrolled – teaching quality, for example – then the analysis is still legitimately open to challenge, however statistically sophisticated it may be.

In conclusion

One of the questions that the review sought to answer was whether we ought to put more weight in results from judgemental methods or statistical methods. Unfortunately, we 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. Now, although this is undoubtedly true, precisely the opposite argument could equally be made. When monitoring comparability, we might be actually 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 between the two is not obvious, which probably recommends using both types of approach wherever possible.

This paper ends by reflecting upon a quotation from an independent panel of experts, who were invited to comment on the maintenance of standards in our A level examinations a few years ago. They concluded that: "There is no scientific way to determine in retrospect whether standards have been maintained" (Baker, *et al*. 2002). At first glance, this might be taken to dispute the very idea of monitoring the comparability of examination standards with any precision.

Well, on the one hand, despite over 50 years of monitoring comparability in England, we still have not reached consensus on exactly what comparability *means*, when two examinations are designed to different content and statistical frameworks. Indeed, even within the measurement profession, there are some who would not accept the legitimacy of certain forms of comparability – such as between subjects or over extended periods of time – while there are others who believe that these forms ought to be prioritised. What hope, then, for a science of comparability monitoring?

On the other hand, given that our field that is premised upon approximation, compromise and (above all) action, I prefer to conclude that: we have seen genuine technological progress in the development of our methods for monitoring comparability; and that our methods do provide at least reasonably defensible insights into the expression of this enigmatic concept, even if our conclusions are necessarily tentative and open to debate.

References

Baker E., Sutherland, S. & McGaw, B. (2002). *Maintaining GCE A level standards: the findings of an independent panel of experts.* London: Qualifications and Curriculum Authority.

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.