COMPARABILITY MONITORING: PROGRESS REPORT

Paul E. Newton[1]

Abstract

This conclusion presents a synthesis of the major themes that have emerged throughout the book. Some tentative answers are offered to the questions that motivated the review and some of the underlying tensions that make research in this field so challenging are revisited. The conclusion ends by considering prospects for the future of comparability monitoring in England.

1 Tentative answers

A number of guiding questions were identified at the beginning of the book. Given the complexity of the issues, and the controversies that still persist, it would be impossible to provide a definitive answer to each one. However, not to provide any answers would be unnecessarily precious since there are conclusions that can reasonably be drawn, albeit somewhat tentatively. Where possible, the conclusions drawn from this review will be contrasted with those drawn during the first and second reviews of comparability monitoring research (Bardell *et al*., 1978; Forrest & Shoesmith, 1985).

1.1 To what extent have the trends in different techniques reflected real methodological progress?

Comparability monitoring has been a prominent feature of examinations research in England for more than 50 years. During this period there has been a substantial expansion in the number of students entered for formal examinations, to a point where almost all students are now examined for the GCSE and more than a third of the cohort is examined at A level. This period has also witnessed a substantial increase in regulation, driven particularly by a concern to ensure comparability. This has been associated with a reduction in the number of examining boards and examination syllabuses, the introduction of common subject cores, criteria and grade descriptions, and the requirement to adhere to codes of practice for processing examinations. Where, once, the examining boards were largely self-regulating, this is now formally the responsibility of an independent regulator. This, in turn, has had implications for the conduct of comparability monitoring research, much more of which is now being initiated and funded by the regulator.

The past half century has not only seen change in the examinations context, it has also seen change in the context of research. Processing national examinations necessarily involves the collation of vast quantities of results data. Mechanisms for dealing with such data have changed beyond recognition, making it possible to store, manipulate and analyse full cohort results easily, rapidly and cheaply. Statistical

techniques have improved radically too, enabling data to be investigated using far more sophisticated analytical models; and the software for running these analyses has not only improved, but has become far more widely available (see also Bell & Greatorex, 2000).

Within this context of change, numerous techniques for monitoring comparability have been developed. The history of judgemental techniques – which rely upon the inspection of performances from different examinations – has witnessed, most notably, the development of ratification, identification and paired comparison methods. Similarly, the history of statistical techniques – which involve comparing results from different examinations whilst controlling for cohort characteristics – has seen the development of common test methods, common examinee methods and multilevel modelling methods.

As new methods have been developed, they have sometimes simply displaced previous ones. Thus, in recent years, the paired comparison approach has displaced ratification and identification methods. Other times, the new methods have merely followed in the wake of previous ones, rather than displaced them. For example, despite a flurry of research in the 1960s and 1970s, which relied upon common test methods, enthusiasm fell substantially during the 1980s, to the extent that their use had largely been discontinued by the mid 1980s (Forrest & Shoesmith, 1985). However, a new wave of enthusiasm for purely statistical methods began in the early 1990s with the application of far more sophisticated statistical modelling (e.g. Tymms & Fitz-Gibbon, 1991). This movement gained momentum during the late 1990s, to the point where multilevel modelling methods are now routinely deployed. Finally, on occasion, newly developed techniques have just failed to gain widespread support, as happened with the distribution method. This raises an important question: to what extent have these trends reflected genuine methodological progress?

Progress in the development of judgemental methods

Judgemental methods have been used since the earliest days of comparability monitoring and their use has continued to the present day. Indeed, judgemental methods have been the mainstay of comparability monitoring research throughout.

From the late 1960s to the 1980s, the choice of judgemental method basically came down to a decision between ratification (where examiners are asked to ratify, or to repudiate, decisions on grade boundary marks) and identification (where examiners are asked to identify grade boundary marks anew). In both of the two major reviews of comparability monitoring efforts the substantial majority of studies involved ratification. Yet, even the second review was undecided on which alternative was preferable, commenting that: 'Identification gives results with less confidence, but repudiation makes it difficult for boards to take corrective action.' (Forrest & Shoesmith, 1985, p. 43). Both of the techniques were recognised to have numerous weaknesses.

The paired comparison technique was introduced with the intention of remedying certain key limitations of earlier methods; most notably, by controlling for severity of individual examiner judgement. Usefully, it also enabled subtle statistical analyses of judgements, providing important insights into the validity of the procedure. It has proved to be fairly simple and flexible, although it does have a tendency to be tedious and time consuming for judges. On balance, it seems fair to conclude that the introduction of paired comparison represented a genuine methodological advance from earlier cross-moderation methods.

The major issue concerning the use of judgemental methods – which still remains unclear despite decades of reliance upon them – is just how accurate judges can be expected to be. As noted in the report of a major investigation into the comparability of standards over time, approaches like ratification and identification assume that senior examiners are able 'to spot a borderline script at twenty paces' (Christie & Forrest, 1980, p. 21) despite those performances being responses to entirely different tasks. Yet it is a commonplace finding of examinations research that quality of performance is highly task-sensitive (e.g. Cresswell & Houston, 1991). How well even senior examiners are able mentally to control for task complexity – and thereby to identify the true levels of attainment that reside beneath qualitatively different performances across examinations – is questionable.[2] There is a body of research from England which suggests (at best) that they may be unable to do so very precisely and (at worst) that their judgements may be susceptible to substantial systematic error (e.g. Good & Cresswell, 1988; Cresswell, 2000; Baird & Dhillon, 2005). Indeed, there is research from further afield which suggests that judges are poor at identifying differential task difficulty, per se, let alone controlling for it (e.g. Impara & Plake, 1988).

Although paired comparison methods do not require judgements of absolute grade-worthiness – merely of relative worth – they still require senior examiners to adjust their perceptions of task performance, to control for differential task difficulty across examinations, and thereby to spot the true levels of attainment that reside beneath qualitatively different performances. And appearances can be deceptive. An examination paper which looks fairly straightforward to a senior examiner may not have been experienced in that way by students. If the senior examiner fails to take this into account when comparing performances between examinations then the judgement of comparability may prove to be inaccurate. Incidentally, even consistency of judgement, both within and between judges, may provide spurious reassurance, since consistency is no guarantee of accuracy. The appearance of comparability may be similarly deceptive to all involved.

These general limitations of judgemental methods affect even the most straightforward of comparability contexts: when comparing different versions of the same examination (e.g. from one year to the next). When entirely different examinations are under comparison – which is typically the case during monitoring exercises – there is the additional complication of judges having to identify a construct, or trait, which is common to both. If senior examiners find themselves simply unable to identify a common basis for comparison (whether directly within

paired comparison or indirectly within ratification and identification) then they will be unable to make any sense of the task with which they are faced. This has proved problematic when comparing standards across decades (e.g. Christie and Forrest, 1980; see also Patrick, 1996) and poses a substantial challenge for the comparison of standards across subjects.

An interesting story remains untold as to exactly why distribution studies were never pursued. The work of Johnson & Cohen (1983) was given short shrift in the second review of comparability studies, which acknowledged that the boards were aware of the problems with identification and ratification, but which concluded that: 'Despite their optimism, Johnson and Cohen were no more successful in solving these sorts of problems than the boards have been.' (Forrest & Shoesmith, 1985, p. 45). Unfortunately, the review gave little indication of exactly why the boards were dissatisfied with the new approach. Until this debate is revisited it will be impossible to determine whether distribution methods represent a real methodological advance over identification or ratification methods, or whether they might even challenge the current supremacy of paired comparison methods.

Progress in the development of statistical methods

Statistical methods for monitoring comparability have also been in existence for a long time, but their use has been limited. During the 1960s and 1970s, with the introduction of a new examination for school leavers around the middle of the ability range (the Certificate of Secondary Education), and with standards to be linked across a large number of new examining boards, the use of common test methods seemed to offer a promising pragmatic solution. From the outset, the inherent weakness of this approach was acknowledged by most researchers: the common test would inevitably measure a somewhat different construct from that assessed by the examinations under comparison; and the more similar it was to any one of those examinations, the more biased it would be in favour of the students who sat it (Newbould & Massey, 1979).

The use of statistical methods is generally assumed to be premised upon a principle of control: differences in average examination results which remain, once the characteristics of respective examination cohorts have been controlled for, are attributed to differences in grading standards between those examinations. From this perspective, the cohort characteristics that need to be controlled for are those related to attainment in the examinations under comparison. So, for example, the use of an aptitude test for monitoring comparability would control for the impact of aptitude upon attainment. The fundamental limitation of the common test method is that it only controls for the impact of factors measured by the test.[3] This is a problem since attainment is affected by a wide range of factors – such as amount of time spent studying, quality of teaching, quality of educational resources, etc. – and, unless all of these factors are effectively controlled for, statistical analyses may generate misleading conclusions.

To draw valid conclusions from statistical methods, where variables remain uncontrolled, it must be assumed that those uncontrolled variables would not have impacted differentially upon attainment in the examinations under comparison. Unfortunately, this assumption can often be problematic. It is not uncommon for examination cohorts to be clustered according to variables which might be expected genuinely to impact upon attainment. One example might be the impact of a national teaching strategy upon the quality of teaching from one year to the next – students who had received better teaching would, in this instance, be clustered in the second cohort. Another example might be the impact of differential motivation; where, for example, students who studied for certain examinations tended to study harder than those who studied for others. Indeed, differences are sometimes designed into examination syllabuses specifically to impact differentially upon attainment, for example, with the intention of making students more motivated (e.g. Jones, 1997).

From the perspective of statistical control, it seems fair to conclude that common examinee methods represent a significant methodological advance over common test methods, since more of the factors which impact on attainment will be controlled for (given that the analyses are limited to a single group of students, all of whom take both examinations). Some of the earliest implementations of common examinee methods involved the use of all candidates who entered for the same examination with different boards. Indeed, Forrest & Shoesmith (1985, p. 9) suggested that this was, in a sense, the paradigm for all comparability studies. It does, however, suffer from the limitation that those students who voluntarily enter for both examinations are unlikely to be representative. If so, then conclusions from results may, once more, prove to be misleading.

Despite this reservation, it is tempting to assume that common examinee methods must be the ultimate in statistical control: surely, if analyses are restricted to a single group of students, then all relevant causal factors must be controlled for? This is incorrect, though, as becomes especially apparent when common examinee methods are used for monitoring comparability between subjects. In this context it seems fairly obvious that students may (both individually and on average) study harder for certain subjects than for others, or be taught better (both individually and on average) in certain subjects than in others.

From the perspective of statistical control, multiple regression methods – and multilevel modelling techniques in particular – seem to offer the potential for the ultimate comparability study. As long as a variable can be measured, either directly or by proxy, it can be fed into a sophisticated multilevel model which is specifically designed to accommodate the kind of cohort clustering that is common in the world of examinations research. In theory, then, the use of these techniques represents a genuine methodological advance over both common test and common examinee methods. Indeed, in a sense, the more sophisticated statistical techniques can be understood potentially to subsume the other methods.

In practice, though, the potential of sophisticated statistical modelling – and multilevel modelling in particular – is unlikely ever to be fully realised for a range of

reasons, of which the following three are particularly significant (see also Chapter 10 of this book; Baird & Jones, 1998).

First, unmeasured factors. If there is even one critical factor missing from a multilevel model, then conclusions may be misleading. For example, it would be quite possible for students to be clustered within examinations according to the quality of teaching experienced. This might occur if, for example, an unusual, although rewarding examination syllabus was chosen far more frequently by enthusiastic teachers. Assuming that teaching quality had a substantial impact on student attainment, but that no measure (or proxy measure) of teaching quality was provided, the conclusions from a monitoring exercise based upon multilevel modelling would be misleading. Multilevel models offer the potential to control for factors that other statistical methods simply cannot control for. However, if the more sophisticated techniques do not actually include measures of these factors then, in practice, they may not represent much of an advance over the less sophisticated ones.

Second, poorly measured factors. This is essentially just an extension of the first point. If a factor is measured, but badly, then it will not be possible to take proper account of it within a monitoring exercise based upon multilevel modelling.

Third, and more subtly, a limited theory of causation. The standard use of multilevel modelling for monitoring comparability seems to be premised upon the following theory of attainment in the examined syllabuses: amount A of factor F will, on average, have impact I upon results in both examinations. According to this theory, it is possible to derive estimates such as the following: if a student studies for ten hours then this will have an impact of 0.045 of a grade on her examination outcome, regardless of whether this is an examination for syllabus x1 (a well-conceived syllabus whose structure facilitates learning of subject x) or for syllabus x2 (a poorly conceived syllabus whose structure hinders learning of subject x). The precise size of this impact is estimated from the modelling of live data, i.e. by modelling the impact of study time upon examination results while all other factors are held constant, at this point in the analysis treating grades from syllabuses x1 and x2 equivalently.

Unfortunately, even if all relevant factors were measured accurately and included in the model, the underlying theory of attainment would compromise the modelling if it turned out not to be true. That is, if it were not actually true that a specific amount of a certain factor had the same impact upon attainment across the examined syllabuses, then the modelling would be undermined. For example, it might be true that ten hours of study following syllabus x1 caused twice the impact upon attainment of ten hours of study following syllabus x2. Yet, this would not be reflected in the modelling, being premised upon a theory of equivalent impact from equivalent input. In theory, the use of interaction terms could enable the investigation of differential impacts for equivalent inputs. However, in practice, it is not clear how the computations could disentangle differences in impact upon attainment across syllabuses from differences in the grading standards of their respective examinations.

The same might be true for other factors which might very well be clustered by examination syllabus. For example, even if a plausible measure of teaching quality could be developed, it might not be true that equivalent teaching quality would have an equivalent impact upon attainment across examined syllabuses. One might simply be harder to teach than another, thereby requiring higher teaching quality for the same attainment gain.

Conceptualising comparability

To achieve substantial methodological progress – indeed, to be able to conclude that substantial methodological progress has been made – it is often important to grapple with the theoretical underpinnings of a field. In the field of comparability monitoring research, there is a requirement to identify with precision what we mean by terms such as demand, difficulty, quality, standards, comparability and so on. It is very easy to use words like these quite loosely or heuristically. But there comes a time when precision is necessary.

Unfortunately, the past half century or so has seen little in the way of focused theoretical analysis. With the exception of some notable contributions (e.g. Christie & Forrest, 1981; Cresswell, 1996; Baird *et al*., 2000; Goldstein & Heath, 2000; Pollitt, this book) few researchers have tackled the major questions of meaning and purpose head-on. While progress has undoubtedly been made with conceptualising comparability, there is clearly still a long way to go. Indeed, as suggested in the commentary by Oates, fully to understand the construct of comparability we need to interrogate it from a wide range of perspectives; not just theoretical, but historical, sociological, political and so on.

1.2 To what extent can each of the methods be improved and is there scope for developing entirely new approaches?

Contributors to this book were invited to consider the scope for improvements to be made to the techniques under review, and numerous avenues for investigation have been suggested. With paired comparison methods, for instance, the obvious next step will be to use scripts from various points along the mark scale of each examination (rather than simply from grade boundary marks) to quantify any degree of difference in grading standards observed. It will be important also to continue to investigate the strengths and weaknesses of the rank-ordering method, since this offers economies of time, making the task less tedious for judges.

With multilevel models, it will be important to continue to develop measures of critical control variables. It will also be important to investigate the strengths and weaknesses of different approaches to modelling, and the inclusion or otherwise of interaction terms.

There is also a case for revisiting certain of the methods which have been identified as particularly significant, but which have seen limited development. The common examinee approach may be a case in point; particularly, its use in monitoring comparability between the same examinations offered by different examining boards

(see also Bell & Greatorex, 2000). The use of this method seems far from unreasonable nowadays, with common subject cores, criteria and grade descriptions which apply to all syllabuses in a subject area. Moreover, by setting up an experimental design in advance of the examination period, it would be possible to engineer far more representative samples than were possible in the original studies of the 1960s. Obviously, there would be an expectation that students would perform better in the examination whose syllabus they actually studied (although, if they did not, then this might raise very significant questions concerning comparability). Yet relative patterns of achievement across the examinations might still provide useful insights into comparability.

It is probably fair to conclude that – given more than 50 years of research using these techniques – there has been insufficient research into their validity and reliability. Such studies are not straightforward to conduct and are less straightforward to interpret, but they are crucial in establishing the defensibility of comparability monitoring exercises. This kind of evaluative research is far from absent in the literature; in fact, it has often been an implicit feature of reports on specific investigations. The introduction of new methods into the field carries a particular obligation to demonstrate reliability and validity, so it is reassuring to see that researchers are taking this responsibility seriously (e.g. Jones & Meadows, 2005; Black & Bramley, in preparation; Black & Bramley, in press). More could be done to extend this body of work; for example, by replicating judgemental exercises using multiple groups of judges, or by answering the same comparability questions using different multilevel models or software packages, or by investigating the stability of multilevel model parameters and variance components using (randomly and non-randomly) split samples, and so on.

Further investigations into the judgement of senior examiners in monitoring exercises would be worthwhile. For example, through robust experimental pre-testing, it would be possible to engineer pairs of parallel examinations to be substantially more or less difficult than each other. These non-equivalent examinations could then be administered to students, and scripts at various mark points selected for scrutiny. If the test construction and linking was done with sufficient care – such that we could be fairly confident in the degree of non-equivalence of the examinations – we would be able to explore the degree of sensitivity of judges to these differences.

1.3 Are certain techniques to be preferred over others?

As noted earlier, there seem to be reasonable grounds to prefer paired comparison methods over identification and ratification techniques. And there seem to be reasonable grounds to prefer multilevel modelling to common test and common examinee methods (particularly if the former is understood to subsume the latter, as noted earlier). But ought we to prefer judgemental methods to statistical ones, or vice versa? Here, the answer is not at all clear. The reports of the first and second reviews of comparability monitoring research concluded as follows:

> cross-moderation involving the boards' examiners (possibly with outsiders too) is the most fruitful and sensitive of the methods available for the study of comparability.
>
> Bardell *et al*. (1978, p. 36)

> the boards believe that cross-moderation is probably the best methodology to pursue at the present time.
>
> Forrest & Shoesmith (1985, p. 45)

Until recently, neither the boards nor the regulator would have placed a great deal of stock in statistical methods for monitoring comparability, given that even the best of the methods left certain key variables uncontrolled. Nowadays, with the widespread adoption of multilevel modelling, there is a renewed interest in purely statistical approaches. Multilevel modelling methods are not without their weaknesses, but clearly nor are any of the judgemental methods.

It is interesting to speculate upon why judgemental techniques have been so influential in the history of comparability monitoring, while statistical techniques have been far more frequently and resolutely dismissed, particularly by the examining boards. Perhaps it is because simple statistical techniques are so easy to criticise from first principles, given their inability to control for all relevant factors (while the limitations of judgement are far less easy to identify and to characterise). Or perhaps society is simply more willing to trust the judgement of senior examiners (over professional statisticians) in this context?

The primary argument in favour of judgemental methods was expressed in the first review as follows:

> Perhaps one of the greatest advantages enjoyed by cross-moderation over other comparability methodologies is that it comes closest to a simulation of the normal task of an examining board in its moderation and awarding procedures. It relies on the actual examination scripts rather than, for example, the results of general ability or specially constructed subject tests. It also invariably utilises the very examiners who are responsible within the boards for the syllabuses, examination papers and forms of assessment.
>
> Bardell *et al*. (1978, p. 30)

On the other hand, precisely the opposite argument could be made: that methods for monitoring comparability ought not to mirror methods used for achieving it. If there are biases of judgement which might compromise decisions during awarding meetings, then similar ones are likely to bias monitoring exercises, and to similar effect. Perhaps alternative insights into comparability are best provided by using alternative methods? The first review went on to propose that:

> In addition, cross-moderation has the marked advantage that of all the formal methodologies available for studying comparability it alone depends on human judgment which can allow for variation in intention between examinations; objective monitoring techniques are invariably insensitive to such variations.
>
> Bardell *et al*. (1978, p. 30)

Precisely what 'variation in intention' means is somewhat unclear, but probably refers to differences in demands between examinations. Importantly, though, the assertion that human judgement 'can' allow for this kind of variation remains

debateable to the present day. Even the first review acknowledged that human judgement is highly subjective, which can lead to findings that may not be very reliable.

Given the potential of sophisticated regression models to control for far more of the critical variables, it is not so easy to dismiss the use of purely statistical techniques such as multilevel modelling. On the other hand, given that even these methods are far from infallible, it is not clear that there is a great deal of evidence to opt for them over, say, paired comparison methods. Having said that, given the kind of experiments suggested in the last paragraph of the previous sections, it should be possible to pit multilevel model against paired comparison directly (at least, under certain conceptions of comparability, a point that will become clearer shortly).

1.4 To what extent are the techniques ultimately based upon the same problematic assumptions?

Given the analysis of the preceding sections, the following broad conclusions might be drawn:

1. Judgemental techniques are based on the assumption that senior examiners are able to adjust their perceptions of task performance, to control for differential task difficulty across examinations, and thereby to spot the true levels of attainment that reside beneath qualitatively different performances.

2. Statistical techniques are based on the assumption that it is possible effectively to control for the various factors that lead to attainment in examined syllabuses, and thereby to interpret differences in results between examination cohorts (which remain once relevant factors have been controlled for) in terms of differential grading standards.

In this sense, it is true to say that all statistical methods share a basic common underlying assumption and that all judgemental methods also share a basic common underlying assumption. Whether these underlying assumptions are problematic, though, is a slightly different matter. As far as judgemental techniques are concerned, the assumption might be more problematic for identification and ratification than for paired comparison (since the latter only requires relative judgements of worth). However, it is fair to conclude that the assumption is problematic for all judgemental techniques. As far as statistical techniques are concerned, the assumption might be more problematic for common test methods than for common examinee or multilevel modelling methods (since the latter can control for more factors). However, once again, it is fair to conclude that the assumption is problematic for all statistical techniques. There are no perfect methods for monitoring comparability.

2 Underlying tensions

An evaluation of techniques for monitoring the comparability of examination standards would be complicated enough if the concept of comparability itself were not contested. Unfortunately, it is contested, and this makes evaluation more complicated still. The following section summarises the challenge of alternative

conceptions – which has been considered from a variety of different perspectives by Baird, Coe, Murphy, Newton, Bramley and other contributors to this book – and explores the implications of alternative conceptions for the conduct of comparability monitoring exercises. It proposes that the primary reason for differences of opinion between stakeholders is their different perceptions about how examination results are, or ought to be, used. Since different inferences are drawn from results when they are used for different purposes, different users and stakeholders will prioritise different conceptions of comparability.

2.1 Alternative conceptions of comparability

The relationship between uses of examination results and conceptions of examination comparability can be illustrated using the following three assessment purposes (from Newton, 2007):

1. qualification – individual results are used to judge whether a person is sufficiently qualified for a job, course of instruction or role in life, i.e. whether or not they are equipped to succeed in it

2. selection – individual results are used to predict which applicants – all of whom might, in principle, be sufficiently qualified – will be most successful in a job or course of instruction

3. programme evaluation – aggregated results are used to evaluate the success of educational programmes or initiatives, nationally or locally.

In each case, the critical issue is the inference that is drawn from examination results to support the particular purpose. All sorts of inferences are possible, but some are more legitimate than others. Given the uses identified above, the following inferences might be drawn:

1. a qualification inference – a student with a grade C in GCSE ICT has the essential knowledge, skills and understanding that will enable him or her to operate confidently, effectively and independently in life and at work

2. a selection inference – a student with an A level grade profile of B (English), B (media studies), B (biology) will be more successful in an undergraduate psychology course than a student with a profile of C (French), C (economics), C (chemistry)

3. a programme evaluation inference – a cohort with an average point score of 6.5 for GCSE physics (treatment 1) will, on average, have attained a superior level of knowledge, skill and understanding of physics than a cohort with an average point score of 5.8 (treatment 2).

Each of these different uses might encourage a slightly different perspective on comparability. The qualification inference seems to take examination results as indicators of sufficient competence to function effectively as citizens; in this example, competence in ICT. For the qualification inference to be valid, equivalent grades from different ICT examinations must represent the same capacity to overcome the ICT

challenges of everyday life and work. We might call this a 'contextualised' attainment-based conception of comparability: it is grounded in the attainment of a specific body of knowledge, skill and understanding, but seen in the context of how this attainment is likely to be deployed (such that the degree of proficiency required for the award of a particular grade would need to increase over time as the technological demands of everyday life and work increased).

By contrast, the selection inference seems to take examination results as indicators of potential, or aptitude, for success in any of a range of higher education courses; in this example, psychology. For the selection inference to be valid, equivalent grades from different examinations must represent the same level of general ability. We might call this an aptitude-based conception: it is not grounded in the attainment of a specific body of knowledge, skill and understanding, but in the general ability of students awarded each grade.

Finally, the programme evaluation inference seems to take examination results straightforwardly as indicators of attainment; in this example, attainment in physics. For the programme evaluation inference to be valid, equivalent grades from different examinations must represent the same level of attainment in physics. We might call this a 'straightforward' attainment-based conception of comparability: it is grounded in the attainment of a specific body of knowledge, skill and understanding, and is blind to any contextual factors (such that the degree of proficiency required for the award of a particular grade would need to remain constant across examinations, period).

These are just three possible uses of results alongside three possible inferences from results to support those uses. None of these uses, nor inferences, is necessarily the correct one. Indeed, the point of the example is simply to illustrate that – just as it is possible to identify different uses of results – so is it possible to identify different conceptions of comparability. Other uses or inferences might encourage entirely new conceptions.

The problem generated by alternative conceptions of comparability is that they cannot all be operationalised simultaneously. Consider, for example, the maintenance of standards across a ten year period of time during which teaching quality improved substantially (due to the impact of national teaching initiatives) and during which the ICT challenges of everyday life and work increased substantially (due to the progressive creep of technology into new aspects of life). How ought standards in GCSE ICT to be maintained over time? From the perspective of a straightforward attainment-based conception: a demonstration of the same knowledge, skill and understanding, from one decade to the next, ought to be rewarded with the same grade, regardless of educational or societal change. Conversely, from the perspective of an aptitude-based conception: the increase in knowledge, skill and understanding attributable to improvement in teaching quality ought not to be reflected in grades (where the increase in attainment is not associated with an increase in aptitude), so – to maintain standards under this conception – students would need to demonstrate higher levels of knowledge, skill and understanding over time for the award of the

same grade. Similarly, from the perspective of a contextualised attainment-based conception: as the demands of everyday life increase, so too should the requirement to demonstrate higher levels of knowledge, skill and understanding for the award of the same grade. This fundamental tension plays out in a range of arenas; most notably, during grade awarding, when debating the appropriate uses of results and when monitoring comparability.

Grade awarding

The procedure by which grade boundaries are set for GCSE and A level examinations prioritises the maintenance of standards from one year to the next, in the same (and similar) syllabus(es), within respective examining boards. There is often at least some evidence on broader concerns, such as comparability with parallel syllabuses from other boards. However, as the most recent code of practice makes clear: 'The prime objectives are the maintenance of grade standards over time and across different specifications within a qualification type.' (QCA, 2007, 6.2).

On the one hand, the code of practice is relatively silent concerning the specific conception of comparability which ought to be adopted during awarding meetings. On the other – given that the mechanism for deriving grade boundary recommendations involves human judgement of performances from successive examinations – this might be taken to imply, or at least to default to, a straightforward attainment-based conception.

The main problem with defaulting to this conception is that it leaves the very idea of comparability between subjects quite ambiguous. Indeed, the most recent code of practice says virtually nothing on the matter; nor does it require that any action be taken to ensure it. Instead, the view of the regulator is that a general sense of comparability between subjects is engineered at a much earlier stage, through the specification of subject criteria. This involves the identification of subject content of an appropriate level and the development of appropriately pitched grade descriptions. This is a far looser understanding of comparability between subjects than some would recommend (see Dearing, 1996, for instance).

In fact, the relative silence of the code of practice could be read to leave open the possibility of embracing different definitions in different circumstances, or of acknowledging multiple definitions simultaneously through a process of compromise (see also Baird *et al*., 2000; Baird, this book; Newton, 2005a).

Debating the appropriate uses of results

To some, even among the assessment profession, there seems to be only one possible conception of comparability: the straightforward attainment-based one (see the response from Murphy to Coe *et al*., in this book). Others, though, are more liberal in their views and accept the potential validity of multiple definitions (see the chapter and commentary by Coe, for instance, as well as the chapter by Baird). In fact, a wide range of alternative conceptions have found expression in the various contributions

to this book. None of these can necessarily be elevated as the correct one, nor even as the preferred one.

Although accepting that a straightforward attainment-based conception might be appropriate for grade awarding meetings, Coe suggested that grades might be recalibrated *post hoc* for use in performance tables, or even for selection purposes. As demonstrated by Lamprianou, though, this kind of technical solution can flounder in the face of public confusion and mistrust. This raises a stark challenge. In many countries which share a similar examination system to England there is an undercurrent of suspicion that standards are not comparable between subjects. However, when technical 'solutions' are implemented in response to such concerns, the outcomes may still fail the test of public confidence. Even though recalibration may seem attractive (to some) from a technical perspective – as an attempt to operationalise multiple conceptions of comparability in response to multiple uses of results – the ultimate proof of the pudding is in the eating. And the 'apples and pears' of between-subject comparability often prove unpalatable howsoever they are prepared!

Monitoring comparability

The implications of alternative conceptions of comparability for monitoring exercises are complicated, as the controversies identified in various chapters have indicated. In exactly the same way as for grade awarding meetings, comparability monitoring exercises have tended to be unclear concerning the conception of comparability operationalised. However, since most of the studies have investigated comparability between parallel examinations offered by different boards, and since most have relied upon the judgement of senior examiners, they have probably also tended to default to a straightforward attainment-based conception.

This has not always been true though, and investigations into the comparability of standards over long periods of time have explicitly thrown up questions concerning the appropriateness of alternative conceptions (e.g. Christie & Forrest, 1980; 1981). There is an interesting point to be made here. When monitoring between-board comparability, differences between their respective syllabuses and between the social and educational contexts of their delivery, are often small enough not to have to call into question the appropriateness of a straightforward attainment-based conception of comparability. However, when those differences exceed a certain threshold, this conception is thrown into relief and the monitoring exercise itself may founder. For example, one examiner, during the Christie and Forrest (1980) investigation into standards in chemistry between 1963 and 1973 described the monitoring exercise as: 'A puzzling task, and not very satisfying.' More generally, the researchers concluded:

> What has been established is that 1973 examiners are much more favourably impressed by the efforts of 1963 candidates than were the 1963 examiners. But these efforts are in the context of 1963 Chemistry and it a moot point as to whether the 1973 examiners were able to adjust their mental set to take account of the intervening change in what constitutes an education in Chemistry. Is it not more probable that the 1973 examiners saw these performances in a more favourable light because, in the context of a different approach to

> the teaching of Chemistry, work of this type will be produced by fewer candidates? The pass marks offered by groups 3 and 4 suggest that this is a highly plausible explanation of the pattern of results.
>
> Christie & Forrest (1980, pp. 59–60)

Technically speaking, a straightforward attainment-based conception of comparability is only entirely valid when the examinations under comparison correspond to exactly the same syllabus. This is often the case during awarding meetings. For comparability monitoring exercises, though, this is almost never true. Almost all require some kind of 'mental set' adjustment for differences in syllabus content, and sometimes also for educational and social context.

This is, perhaps, the central conundrum for comparability monitoring: the conception of comparability that is typically assumed to underpin the *maintenance* of examination standards – the straightforward attainment-based conception – may appear manifestly inappropriate for the contexts in which the maintenance of examination standards is typically *monitored*. Yet, to adopt an alternative conception purely for the purposes of a monitoring exercise would seem faintly odd; akin to moving the goal-posts. In this sense, Murphy is surely correct to argue that only a straightforward attainment-based conception of comparability is defensible.

Yet Coe is surely also correct to argue the opposite: multiple conceptions of comparability are required for multiple uses of results. For Coe, comparability needs to be monitored from a range of alternative perspectives, which correspond to the range of different conceptions of comparability. This is not a matter of using multiple methods to achieve a triangulation of findings. On the contrary, it might involve acknowledging that even a single set of results, from a single method, may have more than one valid interpretation, depending upon the conception of comparability in mind.

This brings us back to the earlier discussion of progress in methods for monitoring comparability, where it was argued that common examinee and multilevel modelling methods represent a substantial advance over common test methods since they control for a greater number of the factors that impact upon attainment. It should now be evident that this is only necessarily true in relation to a straightforward attainment-based conception of comparability. In fact, if an aptitude-based conception of comparability were to be adopted for monitoring purposes, then a robust test of general ability would be the appropriate tool for the job. This would be true regardless of the (potentially different) relationships between the reference test and the examinations under comparison. Likewise, a complex multilevel model which measured many background variables would be an inappropriate tool for monitoring comparability under an aptitude-based conception. And it is hard to see how any judgemental approach could have much legitimacy; unless, that is, we are prepared to accept that judges may be able accurately to spot true levels of aptitude (that reside beneath true levels of attainment) that reside beneath qualitatively different performances.

Which is the correct conception to use when monitoring comparability? Given that results are used for many different purposes, any attempt to rule certain conceptions in or out would inevitably prove to be controversial. Typically, though, certain conceptions will often be ruled in or out, albeit implicitly, by the choice of method and the style of interpretation. Transparency could be improved, in these circumstances, by making the reasoning explicit. More generally, consideration ought to be given to the range of alternative conceptions, each time a comparability monitoring exercise is undertaken, even if certain conceptions are subsequently rejected. These deliberations ought also to be included in the report of any exercise.

2.2 Additional issues

Before considering the future of comparability monitoring it is worth underlining a number of the key challenges that investigators face when conducting comparability monitoring research.

Interpreting results

As explained above, we are not yet in a position to defend the claim that either judgemental or statistical approaches are superior and should be favoured. Until there is a weight of evidence which clearly favours one technique or class of techniques over another, the questions of which and how many methods to employ will always arise. Given the inevitable limitations of both statistical and judgemental methods, it might seem sensible, wherever feasible, to conduct both. There is an important rider, though: certain techniques might be more or less appropriate for investigating different conceptions of comparability. This possibility needs to be confronted explicitly at the outset of any investigation.

This highlights a two-dimensional challenge to the correct interpretation of findings from comparability monitoring studies:

1. Different conclusions might be drawn from the findings of different methods (within a single conception).

2. Different conclusions might be drawn from the findings of a single method (between different conceptions).

Separating these dimensions of interpretation can be awkward, but is necessary all the same, to interpret findings appropriately. Within a single conception, we might hope that conclusions from all methods employed would point in a single direction (were it not for the inevitability of error). Where findings do converge, this fosters some confidence in the defensibility of conclusions. Where findings diverge, this might recommend returning a verdict of unproven.

Between conceptions, the patterns of similarity and difference do not have the same significance. Here, we would not necessarily expect findings to converge. Indeed, we might expect them to diverge. Standards might well be aligned under a straightforward attainment-based interpretation, but not aligned under an aptitude-based interpretation.

Acting on results

Even within a single conception, a convergence of findings can prove problematic to act upon. If, for example, an apparent discontinuity over time was revealed for one examining board, then bringing it back into line might simultaneously threaten comparability of standards between boards. This suggests that some consideration would need to be given to prioritising comparability in different planes. It might, for example, be considered appropriate to tolerate an apparent discontinuity of standards over time, within a board, if there was no evidence of discontinuity of standards between boards for the current examinations. However, if a discontinuity of standards were detected between boards – and between-board comparability were to be prioritised – then this would recommend remedial action. This action would be taken in the subsequent examination session or perhaps, for a large discontinuity, over a number of subsequent sessions. Inevitably, though, this would introduce a discontinuity of standards within the board in question which might, at the very least, present a public relations challenge.

Where only a single method has been used the problems of acting upon results are more stark still. Although there is little in the way of systematic research exploring the accuracy of monitoring methods, there is a general feeling that none can be assumed to be very accurate (see, for example, the contributions from Adams, Murphy and Pinot de Moira). Moreover, despite more than 50 years of research in the field, we have still not entirely cracked the problem of quantifying apparent differences in standards.

Finally, even if findings from multiple studies could be interpreted as strong evidence of a difference in grading standards between boards, and even if the size of that difference could be quantified, there is still a challenge in acting upon those results. This is how to decide which of the divergent examining boards, if any, could be said to represent the 'correct' standard.

Communicating results

In the Foreword to the first major review of comparability monitoring studies, the convenor of the GCE Board Secretaries wrote:

> In presenting this booklet to the public, and inviting applications for the full reports, where available, for those who wish to probe deeper, we in the GCE boards have found ourselves in a dilemma. If we merely state that comparability exercises are regularly conducted and do not show our hand, we appear to have something to hide. If we try to explain them, their complexities and limitations invite misunderstanding and misrepresentation. On balance, the preferable alternative seemed to be to 'publish and be damned'. We have, and probably shall be.
>
> Bardell *et al*. (1978, p. 6)

This concern resonates to the present day (see Newton, 2005b; 2005c). Yet, despite a fear that outcomes may be misunderstood and misrepresented, examining boards,

advisory councils and regulators have regularly published the outcomes of comparability monitoring research for the best part of 40 years now. Although this does result in the occasional public damning, the system is surely better for this level of openness.

As was clear from Baird's chapter, though, not only do different stakeholders hold different conceptions of comparability, some hold technically indefensible views; sometimes very adamantly, passionately and loudly (albeit, perhaps, with entirely noble intent). In this context, it is crucial for the results of comparability monitoring exercises to be presented in a careful and considered manner. The findings must be explained as straightforwardly as possible – particular care must be taken with findings from sophisticated statistical analyses such as multilevel models – and any limitations and caveats must be made explicit and highlighted.

3 Future challenges

The nature of public examining in England has remained fairly constant over the past 50 years or so. There have been some major innovations – such as the introduction of differentiated assessment at GCSE and the modularisation of A level – but within most of the traditional subject areas the kind of examinations sat and the methods of marking and grading have remained fairly similar. Times change, though, and an emerging agenda of issues may change the face of examining in England. What implications might these changes hold for the monitoring of examination standards? Here are a few preliminary thoughts.

3.1 E-assessment

E-assessment is becoming a reality of large-scale educational assessment internationally. Although examinations in England are still largely paper-based, procedures for data capture and script marking are increasingly becoming automated. Although e-assessment offers the potential for radically rethinking the nature of examining, it also makes certain traditional approaches seem more attractive than they otherwise might have. This is certainly true with respect to selected-response tests. Although these have always had the advantages of speed, automation of marking and data collection, and reliability, their delivery via electronic media now offers further benefits in terms of the potential for the calibration (and continuous re-calibration) of items within item banks and the potential for adaptive testing. These advantages may even prove to be possible with constructed-response tests.

Although reaping the full potential of e-assessment presents many a technical challenge, if these kinds of examination were to come to predominate in England, the implications for both grade awarding and comparability monitoring would be radical. The likelihood is that procedures would become more experimentally and statistically driven. They would likely become more automated and more detached from human interrogation. Although this might have many benefits, there are inherent dangers here. The conceptual tensions and challenges would not go away, but they might well become substantially obscured.

Even if the nature of public examining were not to change radically in the short- to medium-term, there are ways in which e-assessment can support conventional approaches. For example, presenting performance evidence on-line can overcome cost constraints, allowing the involvement of many more experienced examiners in judging standards. This is likely to improve the reliability of comparability monitoring. The challenge to overcome is how to present large quantities of performance evidence – via the Internet – in a manner which is conducive to the kind of judgements required within comparability monitoring exercises.

3.2 Unitisation

In the past, even though grade boundary decisions were made on a component-by-component basis, the derivation of subject grade boundary marks represented the ultimate goal. This process was independent of the derivation of component grade boundaries: it was based upon information from component-level decisions, but was not entirely driven by it. Examiners within awarding meetings were able to see work from all components (although this did not necessarily involve coursework for practical reasons) and they were able to trade-off decisions on component standards when reaching recommendations on the subject standard. Only subject grades were officially reported to students. In short, from an historical perspective, standards were understood to reside firmly at the subject level. This is still largely true for GCSE examinations.

Nowadays, following the modularisation of A level, this is no longer necessarily true. Standards are judged at the unit level, against unit level script archives, and subject grade boundaries, based upon scaled uniform marks, are pre-set. Results are officially reported at the unit level. At least in practice, if not in theory, A level standards now increasingly reside at the unit level.

If standards within modularised qualifications are increasingly perceived to reside at the unit level then this has obvious implications for comparability monitoring studies. These would need to be focused at the unit level, with implications drawn at the subject level only where appropriate (subject-level conclusions are unclear when students within the 'same cohort' will have taken different combinations of units).

3.3 Selection tests

There has been increasing debate in England over the use of selection tests, particularly for entrance to higher education (e.g. Admissions to Higher Education Steering Group, 2004). To date, the concern has largely been restricted to elite departments within elite universities, whose selectors are forced to make decisions between students whose A level grade profiles are insufficiently distinct (due to ceiling effects). However, if the move toward the use of selection tests increased, this would have implications for the uses of A level results, and consequential implications for monitoring comparability.

Certainly, if selection tests were routinely used in addition to A level results, this might take away some of the pressure to ensure precise comparability between

examinations in different subject areas. Instead, the selection tests would bear the brunt of the burden of indicating aptitude. Of course, the pressure would not be eliminated entirely, as long as results were still aggregated across subject profiles for inclusion in local, regional and national performance tables.

3.4 Functional skills

Functional skills qualifications are presently being developed in the areas of English, mathematics and ICT. Passing a functional skills examination is intended to certify that students have the essential knowledge, skills and understanding that will enable them to operate confidently, effectively and independently in life and at work. As explained earlier, this means that they are intended to fulfil a 'qualification' function: students who pass are deemed minimally qualified for the challenges that life will present in the domains of English, mathematics and ICT.

As noted earlier, this implies a contextualised attainment-based conception of comparability, which is not defined purely in terms of what a student knows and can do. Instead, it ought to be defined in terms of whether what a student knows and can do equips her sufficiently for the challenges of life. If, for example, the challenges of life were to increase, correspondingly more should be expected of students who pass. The difference between this and a straightforward attainment-based conception is subtle, but crucial; especially when considering comparability over extended periods of time.

3.5 Diplomas

One of the most radical changes on the horizon is the introduction of new Diplomas. These will be portmanteau qualifications – comprising combinations of separate qualifications – and will be offered initially in one of five vocationally-oriented lines of learning: society, health and development; engineering; creative and media; construction and the built environment; and information technology. Each Diploma will comprise functional skills qualifications, a qualification covering principal learning in the sector, a project-based qualification and additional qualifications of the student's choice.

The challenges involved in introducing these new diplomas are not dissimilar to those associated with the introduction of the Certificate of Secondary Education in the 1960s. Whereas standards in the CSE were monitored using reference tests (e.g. Skurnik & Hall, 1969; Willmott, 1977) it seems likely that multilevel models would be the preferred statistical tool for the Diplomas. However, this might depend on what conception of comparability was being aspired to. Across such diverse qualifications, perhaps an aptitude-based conception might be deemed defensible, after all? There is a debate to be had here – potentially a heated one, if previous experience is to be a guide.

Conceivably, though, new methods for monitoring comparability might be developed, better to represent alternative conceptions. For instance, it might be worth revisiting the 'fitness-for-purpose' method developed in the mid-1990s by Coles and

Matthews (1998). This method explored comparability between qualitatively different types of qualification (general versus vocational science qualifications), seeking views of expert groups (employers and academic selectors of science students) on the suitability of those qualifications as preparation for the 'next step' in the students' careers. This logic resonates with the contextualised attainment-based conception of comparability noted earlier: comparability concerns whether the level of knowledge, skill and understanding that constitutes the passing standard for each Diploma equally equips learners for progression in their respective fields.

3.6 Provision and regulation

Over the decades, there have been repeated calls for a reduction in the number of examining boards, to reduce the threat of different standards between them. There is some logic in this sentiment, since the more boards offer the same examination the more likelihood that differences in standards will exist. On the other hand, even with a single examining board, and a single examination for each subject, the challenges of comparability would still persist (particularly over time and between subjects). In addition to offering schools some choice over syllabuses, the diversity of boards is beneficial from the perspective of innovation, not simply in terms of delivering and processing examinations but also in terms of the development of techniques for monitoring comparability.

A key issue for coming years will be where the locus of responsibility for monitoring comparability ought to lie. In recent years, although the examining boards have continued to conduct comparability monitoring research, the regulator has increasingly initiated and funded similar work (sometimes contracting specialist researchers from the examining boards to conduct this research on its behalf). Whether this trend will, or ought to, continue is a moot point. Is there a particular role for the regulator in initiating and funding comparability research (as opposed to validity and reliability research) given that this often requires the co-ordination of efforts across examining boards (which would not necessarily be true in relation to reliability or validity)?

3.7 Internationalisation

Finally, the internationalisation of the qualifications market may need to be recognised more explicitly than it has been in the past. Comparability monitoring work has, almost exclusively to date, been concerned with comparability between examinations offered by the English examining boards; especially, within A level and within GCSE. However, with increasing population mobility, qualifications need to have a currency that can cross national borders. A case could be made for extending comparability monitoring work accordingly.

4 In conclusion

> There is no scientific way to determine in retrospect whether standards have been maintained.
>
> Baker *et al*. (2002)

This book has reviewed the development and evolution of the science of comparability monitoring in England. Is it a story of success or of failure? On the one hand, despite more than 50 years of monitoring, there is still no consensus on exactly what comparability might mean when two examinations are designed to different content and statistical frameworks. Even within the assessment 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, we have made substantial progress in the specification of alternative conceptions of comparability and in the development of techniques for monitoring comparability. It may still not be possible to derive definitive conclusions from the application of any particular method, given any particular conception, but that does not mean that findings from comparability monitoring work are either indefensible or not useful. Far from it. Monitoring exercises have the potential to provide valuable insights and a legitimate basis for action in an inherently pragmatic world.

The research described in this book shows how – with appropriate caution and with due reflection on meaning – systematic, *post hoc*, investigations can be designed and undertaken, with the potential to offer persuasive findings concerning the comparability of examination standards. We have developed substantially better methods than existed 50 years ago and are becoming progressively better at interpreting their findings. There is certainly much work still to be done, but substantial progress has been achieved so far.

Endnotes

1. This concluding chapter was prepared by the Lead Editor, with support and guidance from the other members of the Editorial Board, and with advice from a number of chapter authors.

2. This statement is premised upon the assumption that comparability, like assessment more generally, relates ultimately to the quality of the student that produces a piece of work, not to the quality of the work, per se (certificates are awarded to students, not to their work). In this particular case, the quality of the student is being defined in terms of an underlying level of attainment.

3. There is one special case in which controlling for a single factor might be considered sufficient: when the construct measured by all examinations under comparison is identical with the construct measured by the common test. This amounts to controlling for attainment directly, rather than controlling for the many factors that underlie attainment.

References

Admissions to Higher Education Steering Group. (2004). *Fair admissions to higher education: Recommendations for good practice*. Nottingham: Department for Education and Skills.

Baird, J., Cresswell, M.J., & Newton, P. (2000). Would the *real* gold standard please step forward? *Research Papers in Education*, *15*, 213-229.

Baird, J., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: Valid, but inexact*. Internal Report RPA 05 JB RP 077. Guildford: Assessment and Qualifications Alliance.

Baird, J., & Jones, B.E. (1998). *Statistical analyses of examination standards: Better measures of the unquantifiable?* Research Report RAC/780. Assessment and Qualifications Alliance.

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.

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.

Bell, J.F., & Greatorex, J. (2000). *A review of research into levels, profiles and comparability*. London: Qualifications and Curriculum Authority.

Black, B., & Bramley, T. (in press). Investigating a judgmental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*.

Black, B., & Bramley, T. (in preparation). *Using expert judgment to link mark scales on different tiers of a GCSE English examination: A rank ordering method*.

Christie, T., & Forrest, G.M. (1980). *Standards at GCE A Level: 1963 and 1973*. Schools Council Research Studies. London: Macmillan Education.

Christie, T., & Forrest, G.M. (1981). *Defining public examination standards*. Schools Council Research Studies. London: Macmillan Education.

Coles, M., & Matthews, A. (1998). *Comparing qualifications – Fitness for purpose. Methodology paper*. London: Qualifications and Curriculum Authority.

Cresswell, M.J. (1996). Defining, setting and maintain ing standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). Chichester: JohnWiley and Sons.

Cresswell, M.J. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein & A. Heath (Eds.), *Educational Standards* (pp. 69-104). Oxford: Oxford University Press for The British Academy.

Cresswell, M.J., & Houston, J.G. (1991). Assessment of the National Curriculum -

Some fundamental considerations. *Educational Review*, *43*(1), 63–78.

Dearing, R. (1996). *Review of qualifications for 16-19 year olds*. London: School Curriculum and Assessment Authority.

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.

Goldstein, H., & Heath, A. (Eds.). (2000). *Educational standards.* Oxford: Oxford University Press for The British Academy.

Good, F.J., & Cresswell, M.J., (1988). Grade awarding judgments in differentiated examinations. *British Educational Research Journal*, *14*, 263-281.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement*, *35*, 69-81.

Johnson, S., & Cohen, L. (1983). *Investigating grade comparability through cross-moderation*. London: Schools Council.

Jones, B.E. (1997). Comparing examination standards: Is a purely statistical approach adequate? *Assessment in Education*, *4*, 249-263.

Jones, B.E., & Meadows, M. (2005). *A replicated comparability study in GCSE religious studies*. Manchester: Assessment and Qualifications Alliance.

Newbould, C.A., & Massey, A.J. (1979). *Comparability using a common element.* Occasional Publication 7. Cambridge: Test Development and Research Unit.

Newton, P.E. (2005a). Examination standards and the limits of linking. *Assessment in Education*, *12*, 105-123.

Newton, P.E. (2005b). Threats to the professional understanding of assessment error. *Journal of Education Policy*, *20*, 457-483.

Newton, P.E. (2005c). The public understanding of measurement inaccuracy. *British Educational Research Journal*, *31*, 419-442.

Newton, P.E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education*, *14*, 149-170.

Patrick, H. (1996, September). *Comparing public examination standards over time*. Paper presented at the British Educational Research Association annual conference, Lancaster University.

Qualifications and Curriculum Authority. (2007). *GCSE, GCE, VCE, GNVQ and AEA*

code of practice. London: Qualifications and Curriculum Authority.

Skurnik, L.S. & Hall, J. (1969). *The 1966 CSE monitoring experiment*. Schools Council Working Paper 21. London: Her Majesty's Stationery Office.

Tymms, P.B., & Fitz-Gibbon, C.T. (1991). A comparison of examination boards: A levels. *Oxford Review of Education*, *17*, 17-32.

Willmott, A.S. (1977). *CSE and GCE grading standards: The 1973 comparability study*. Schools Council Research Study. London: Macmillan Education.