8

COMMON TEST METHODS

Roger Murphy

Abstract

Aim

To review the use of common test approaches in comparability research, assessing both the advantages of such an approach and the criticisms that have been levelled against it.

Definitions of comparability

The debate about the usefulness of this method partly hinges on the definition of comparability assumed by those considering its use. The approach fits best with a statistical approach to comparability and least well with a standards referenced approach.

Comparability methods

The common test approach relies upon the examinations being compared having been taken by a sample of students, who have also all taken some other common test. The common test results are used as the basis for comparing the standards of different examinations. This is usually undertaken by plotting regression lines, which attempt to estimate the relationship between common test scores and examination scores.

Strengths and weaknesses

The common test approach is a fairly simple method. It can therefore be easily explained to non-technical individuals who are interested in comparability issues. It is also relatively easy to collect the data needed and draw conclusions. However, it does depend upon the common test having a strong and consistent educational and statistical relationship with the examinations being compared. Critics of the approach point to the fact that common tests rarely have anything like the required relationship with examinations of the type for which comparability studies are required.

Conclusion

The method should only be used in circumstances where the relationship between the common test and the examinations to be compared can be studied closely and critically. Any comparability conclusions drawn from such a study need to be interpreted with caution, taking account of the known levels of uncertainty.

1 Introduction

Already there are some clear themes being built up as we work through the chapters of this volume. Comparability in relation to examination standards can be defined and interpreted in quite a few different ways. No one definition stands apart from the others as the best or most appropriate one. However, when it comes to doing very precise things like designing research studies to investigate comparability questions then it is necessary to work with specific definitions, which can be reflected in the design of such studies. Against that background there has been a sustained and earnest search to develop appropriate methods for researching comparability.

The several different approaches, which have been used over the years, can be delineated fairly clearly and that allows chapters such as this one to review a set of studies that share common features. In this case we will look at comparability research methods that employ a common test approach. As with other methods this general approach can be used in a variety of ways to suit the needs of different situations and it can be seen to have both obvious strengths and weaknesses. Like travellers marooned in a foreign land we cannot afford to dismiss any possible form of transport that comes along, and because this approach, on the face of it, offers some possibilities when it comes to researching difficult comparability questions we will now explore it to see what it has to offer.

In broad terms comparability research methods can be categorised as judgemental and statistical, and this chapter will look at the sub-section of the statistical approaches that are based upon common test approaches. In essence this approach involves addressing comparability questions through evidence gained from situations where the groups of candidates being compared have all also been assessed through a common test. In many situations this will have involved the administration of a common test for research purposes. However in other situations common elements may have been included within otherwise different examinations, or alternatively test scores collected for other purposes may be accessed as the basis for exploring comparability study comparisons

Some of the issues addressed in this chapter will connect with those considered elsewhere in this volume. There are some obvious recurring themes in the chapters, especially those that look at the various statistical approaches. There are for example some close connections between the 'subject pairs' (see Chapter 9) and 'common test' approaches both of which attempt to utilise independent evidence about candidates' ability/achievement levels in order to help investigate the comparability of their results on different examinations. In addition 'multilevel modelling' (see Chapter 10) is a statistical approach, which can be applied to the kind of data generated through several statistical approaches. Also this, like other chapters, will need to continue to refer back to the issues raised in the first four chapters, concerning the general context within which comparability research has emerged, including the attempts that have been made to unpick the variety of ways in which comparability itself can be defined.

My aim in this chapter is to give the reader an insight into the variety of common test approaches that have been used in comparability research, and to explore their advantages and disadvantages. Inevitably this will involve some reference to statistical issues, but I won't be replicating the very detailed statistical discussion to be found for example in Chapter 10. By the end of the chapter we will have looked at the role played by common test approaches over a period of some 40 years. Those who have already read the earlier chapters in this volume will be well prepared for the challenges ahead and will know just what a complicated business it is to tie comparability down in anything other than a very partial way (Nuttall, 1979). The common test approach has at times been seen as controversial, and as such has been both stoutly defended and strongly attacked. As far as is possible I want to look beyond the polarised positions often adopted in such exchanges and help the reader to gain a balanced view of the issues raised. The aim will be to provide a platform of knowledge and understanding through which the strengths and weaknesses of this approach can be assessed.

2 A simple approach to a complex challenge?

Amongst the approaches used to study the comparability of public examinations, the common test (or reference test) approach is undoubtedly one of the simplest. Given a situation in which a comparison needs to be made – say between two groups of students who have taken the same subject exam in different years, or the same subject exam through two different awarding bodies, or different subject exams – the most straightforward version of this method is where both groups of students are given an additional common test for the purposes of researching the comparability of the two exams that are to be compared. In such a situation, the common test results are used to compare the results obtained by the students on the other two examinations.

So, if the overall performance of the two groups of students on the common test is similar and grades obtained by the two groups of students on the two examinations is markedly different, then this is taken as an indication that the two examinations are not comparable in terms of their grading standards. Depending upon the design of the particular comparability study and the nature of the comparisons being made (between years, boards, subjects, etc.) different statistical tests can be applied. These can span comparisons of raw test scores, average scores on each test, least squares regression analysis through to more sophisticated approaches such as multilevel modelling. We will look in more detail at those approaches later in the chapter.

Overall the common test approach is neat and straightforward and fairly easy to explain to users of examination results who may be anxious to know how to compare grades obtained by students in different examinations. It hasn't, however, remained free from controversy, and its relative simplicity has on occasions led some to think that it can be easily misinterpreted as providing very clear answers to what are complex and sometimes highly charged questions about the relative standing of grades obtained from different examinations (Goldstein & Cresswell, 1996; Murphy *et al*., 1996; Newton, 1997; Baird *et al*., 2000).

Looking back over more than four decades of comparability research it is fascinating to see the common test approach come in and out of favour on different occasions. The fact that serious doubts have been raised about its suitability in almost every situation where it has been used is not nearly enough to remove it from a comparability scene, where there is no perfect method that can brush aside all others.

Before getting too immersed in the arguments for and against this approach and the added complications of the many different ways in which it can be applied and analysed, I would like to start by presenting a reasonably straightforward explanation of a simple application of this approach.

3 A simple explanation of the common test approach

To illustrate a relatively straightforward use of the common test approach let us take a simple situation in which two different groups of students have entered for two different GCSE mathematics examinations set by two different awarding bodies (A and B) in the same year. As part of an attempt to investigate whether these two examinations have been graded in an equivalent way, an additional common test is given to both groups of students. This specially constructed common test is a test of mathematical ability that is deemed to be appropriate for GCSE mathematics students.

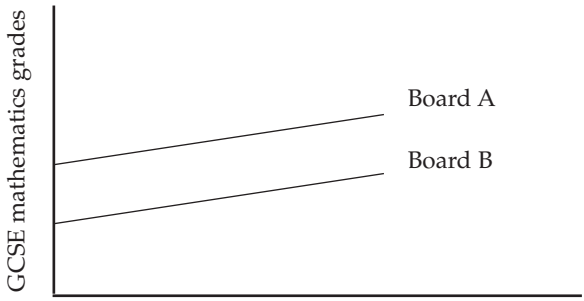
Once the students have taken the common test and it has been marked, there are then two results for each student, their results on the common test and their results on the GCSE mathematics examinations. By comparing the two sets of GCSE and common test results it may be possible to form an impression of how performance in the two GCSE mathematics examinations compares with performance in the common test. If students taking Board A GCSE mathematics obtain higher grades, relative to their common test scores, than those taking Board B, then a case can be made to conclude that the mathematics examination of Board A is easier than that of Board B.

The most common analytical statistical technique used in these situations is linear regression analysis (this is explained more fully in Chapter 10). Using this technique, best fit regression lines can be fitted to the data and plotted to show how in this case the GCSE mathematics grades awarded by Boards A and B relate to the scores obtained by the students on a common test of mathematics ability. Figure 1 shows what the results of such an analysis might look like.

In the majority of similar studies (Bardell *et al*., 1978) the common test is used to predict examination grades on the basis of fitting best fit linear regression lines to the available data. In order to keep things simple, the two regression lines in this case have been shown as running parallel with each other, indicating that the marks and grades for the two GCSE examinations had similar linear relationships, albeit with Board A appearing to issue higher GCSE mathematics grades relative to those issued by Board B.

In terms of the simple level of analysis, this example illustrates a situation in which the results obtained might be used as the basis for arguing that the standards being used by Boards A and B for issuing GCSE mathematics grades might not be the same. Taken at face value, the analysis reveals that students with the same score on the test of mathematical ability would on average get a higher GCSE mathematics grade with Board A than they would get with Board B.

Figure 1 Regression lines for common test GCSE mathematics comparability study

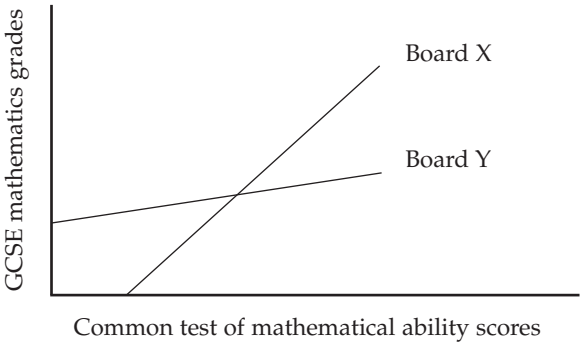


The type of comparison illustrated in Figure 1 could arise either from studies that compare standards between boards (as shown) or that compare standards across subjects. Studies that compare standards across subjects would be more likely to use a general ability reference test than a subject-specific one, but in other respects the design of the study and the representation of the results would be along exactly the same lines as shown in Figure 1. Studies that use common tests to study possible changes in standards over time require different forms of analysis, which we will discuss later.

Newbould & Massey (1979) provide a fairly comprehensive discussion of the statistical issues that can arise from common tests comparability studies. When the results produce parallel regression lines, such as those shown in Figure 1, then it is possible to test for differences between them using analysis of covariance (Brownlee, 1965). However, as Newbould & Massey point out, examination data rarely behave that consistently and less uniform outcomes are frequently encountered. Figure 2 shows another kind of outcome from this type of study. In this case the regression lines are not parallel and intersect. Such an outcome is far from uncommon and may be taken to reflect the fact that the common test may have different relevance for the candidates in the two groups being compared.

This phenomenon is not just a theoretical possibility, as it has arisen regularly in common test studies which check for parallel regression lines (Meadows, 2003; Stringer, 2005). The consequences of this happening are a quite serious threat to the common test approach and in most cases will be seen to limit the conclusions that can be drawn about the comparability of the tests being compared. In the case

Figure 2 Non-parallel regression lines for common test study



illustrated in Figure 2 the only conclusions that can be drawn are that the study has revealed an inconsistent pattern when it comes to comparing the grading standards applied by Boards X and Y. Board X candidates might be deemed to have been graded more leniently than Board Y candidates on the higher grades, but more severely than Board Y candidates on the lower grades. An alternative conclusion is that the relationship between the common test scores and the examination grades awarded by these two boards is so uneven that it is safer to conclude that the method cannot adequately address the issue of the comparability of the two examinations.

An important lesson here is that anyone conducting a common test comparability study of this type should carry out a precautionary test to see if the regression lines produced are reasonably parallel. In the early days of common test comparability studies this was often overlooked (Willmott, 1980), and that is a serious issue as lack of parallelism poses a serious threat to the validity of any conclusions drawn from a study of this kind.

Newbould & Massey (1979) also point out that it is most common in carrying out such analyses to calculate the regression of the examination grades awarded on the scores on the common test. Such an approach can address the question of whether candidates from different boards, for example, with equivalent scores on the common test receive the same mean grade on the examination. Nevertheless it is also possible to do the analysis the other way around by asking whether candidates from two boards, for example with indistinguishable grades, would receive the same mean score on the common test. In such a case one would be looking at the regression of the monitor on the marks awarded in the two examinations. Figure 3 shows what this type of analysis might look like for a single examination, and if applied to more than one examination it would allow comparisons to be made between, for example, common test scores and the marks and grades awarded by different boards. By plotting similar graphs for the examinations to be compared an overall picture will emerge of the nature of the relationship of each examination with the common test. So, for instance, the borderline scores chosen for grade A in two different examinations might be seen to equate with different scores on the common test. In such a case feedback could be given to those involved stating that grade A in Board

H was shown to be equivalent to a score on the common test that was two marks higher than the grade A of Board G.

Figure 3 Regression of common test on examination marks

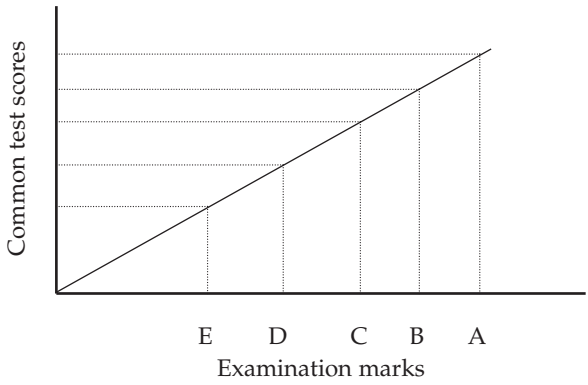


Figure 3 shows a modelling of the relationship between common test scores and examination marks. It also allows estimates to be made of the common test scores equating to each of the minimum examination grade cut-off marks. Newbould & Massey argue that this method leads to a form of feedback that may make more sense to examiners, because the results can be expressed in terms of difference in where grade boundaries have been placed in relation to the total examination marks for a particular examination. In addition they also point out that this approach to analysis typically produces different results from the approach that looks at the regression of the examination results on the common test scores. This represents one more warning that estimates derived from studies of this kind need to be treated with great care and should only be seen as statistical approximations rather than very precise indicators of the relative difficulty of different examinations.

Much of the discussion presented above has assumed straightforward linear relationships between monitor scores and examination results. In reality such relationships are rarely that tidy. Because both examination marks and common test scores are approximations (Murphy, 2004) rather than being highly precise and accurate measures, tolerance limits need to be placed around all such scores. Furthermore, common test comparability studies are frequently conducted on samples that are drawn from larger populations of candidates taking particular examinations, so further account needs to be taken of uncertainty factors associated with drawing conclusions about populations from samples included in research studies.

Figure 4 illustrates this point with findings from Skurnik & Hall (1969), who conducted a common test study of nine 1966 CSE English examinations. Skurnik & Hall, as one part of their study, calculated mean CSE grades and mean scholastic aptitude (CP66) scores for a sample of candidates from each of nine CSE boards. Based upon comparisons between mean CSE English grades and mean common test

scores a regression line was plotted as shown in Figure 4. However, as both the regression line and the board points in this figure are estimates based upon data collected from a sample of schools that agreed to participate in the study, they felt that it was necessary to consider the margins of error that were associated with the regression line and the board points in this figure.

Skurnik & Hall treated the board points as fixed and plotted error probability limits on either side of the best fit regression line, taking account of the fact that both the regression line and the board points were only estimates 'based upon a fraction sample of each board's candidates'. The confidence interval shown therefore takes account of the known standard errors both in the regression line and the points that have been plotted to indicate an approximate position for each CSE board based upon its mean grade and mean common test score. These represent a 95% level of confidence (i.e. plus or minus two standard errors), and define a broad band within which nearly all of the nine boards can be seen.

Figure 4 CSE board comparability study results, Boards 1–9, 1966

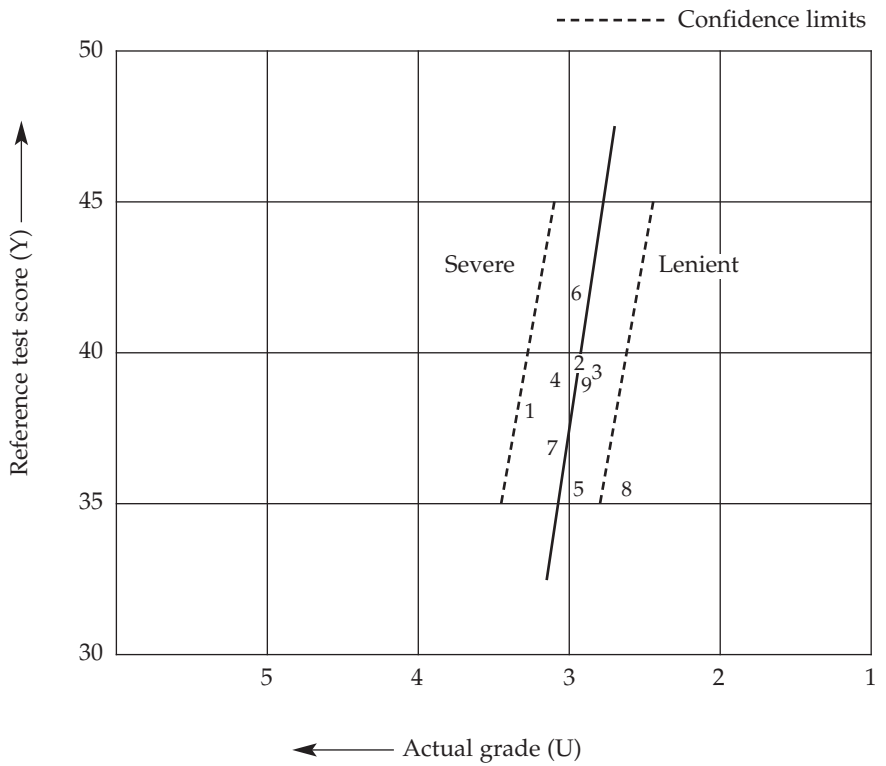


Figure 4 illustrates that with a study of this kind, and with some basic estimation of margins of error, it is not possible to have confidence in the differences in results representing real differences in standards between eight of the nine CSE boards. This is a very graphic representation of the way in which uncertainties concerning the accuracy of common test study results need to be factored into any meaningful interpretation of the results. Without such an appreciation of errors of measurement Figure 4 could have been interpreted quite differently as indicating a good deal of variation say in the grading standards applied by different CSE boards to their English examinations in 1966. Equally one could argue that the error estimates that they have calculated are an underestimate of the real level of uncertainty, given that the common test and exam grades themselves could have been assigned their own 'standard error' estimates, which when combined with the sampling errors could have increased the width of the confidence limits even further.

This single example illustrates a more general point, which we will return to several times. Common test studies of this kind will never give highly accurate outcomes that will allow precise comparisons to be made between boards, subjects and years. Like other statistical approaches to comparability the common test method can provide interesting data, which need to be handled cautiously by those who know both the strengths and weaknesses of the approach. Such an argument needs to be related to wider arguments about the very nature of educational assessments conducted through the UK public examination system, which do not fit the major assumptions of standard psychometric assessments (Gipps, 1994; Broadfoot, 1996; Black, 1998) and which need to be treated accordingly. In the same way that it is foolhardy to expect high levels of precision in the results of such examinations it is even more foolish to expect high levels of precision from research studies that seek to compare different examinations.

Already a section that aimed to present the basic common test approach simply has started to become quite complicated. There is no apology for this because as with all the other statistical and judgemental approaches being discussed in this volume, there are many assumptions being made when one tries to investigate comparability issues through this type of approach. There is as a consequence a good deal more unpicking that we need to do, in later sections, in order to explore the full range of issues involved in evaluating the common test approach.

Before moving on we also need to note at this stage that common test studies that investigate comparability across years generally need to use different forms of analysis. In such studies different groups of students take the common test in different years and their results are analysed year by year using regression analyses to estimate the average examination achievements for students with same scores on the common test. Later in this chapter we will look at an example of such a study, where the results are set out in the form of a graph (Figure 5) showing the grades obtained by students with the same general ability test scores over a period of 16 years. Here the analysis seeks to explore trends over time in the relative performances of students on a common test and A level examinations in a variety of subjects. Even though this approach is somewhat different to the one illustrated

through Figures 1–3, the interpretation of the data still tends to depend upon the same issues and assumptions faced in the other studies.

So having started to grapple with the many issues that can influence the interpretation of common test comparability research studies, we will now turn to consider what is probably the most critical issue of all, which is the nature of the relationship between the common test used in any study and the examinations which it is being used to compare.

4 Different ways of incorporating common tests in both operational examinations and research designs

The focus of this volume is upon the use of methods to research the comparability of GCSE and GCE examinations, rather than exploring the ways in which examination boards attempt to ensure the comparability of examination standards through the approaches they use for setting, marking and grading such examinations. In that sense, the major focus of this chapter is upon situations where common tests are given to students as a means of investigating the comparability of examinations. In a limited number of cases common test results have existed for other reasons and have been accessed as the basis upon which to carry out comparability analyses (Massey *et al*., 2003).

It is, however, also worth noting that some GCSE and GCE examinations have at times had common test elements built into them, usually as a means to help ensure that the examinations lead to comparable grades. In GCSE subjects that use tiered papers, all candidates will normally attempt some common elements of assessment as well as some different elements. The issues arising in tiered GCSE examinations are essentially very similar to those already discussed. A fuller treatment of them can be found in Baird *et al*. (2001). Although the common papers can be used to compare grading standards in optional tiered papers, which are taken by sub-sets of the complete entry, the interpretation of any apparent differences is confused further by the fact that students of different levels of achievement are usually entered for easier or harder optional papers, and in such a situation other possible causal factors can limit the interpretation of apparent differences in the way optional tiered papers are marked and graded. Similarly, multiple choice tests, where used, can be constructed so as to include some common test items as the basis for equating performance from one year to the next. Within National Curriculum testing the 'anchor test' approach is well established and frequently used as one basis for statistical test equating.

Where common test elements are used, whether for researching comparability or for attempting to ensure comparability in live examinations, the same issues arise in terms of the nature of the common test elements and the way in which they relate to the aspects of candidate achievement being assessed through the other elements of the examination. To be useful, common test elements need to relate quite closely to the other assessment elements: otherwise they simply give complementary information about the achievements of the students and have little to offer in relation to predicting candidate performance on the whole examination. Linked to that issue

is the requirement for the common tests to be equally fair for all students and not introduce spurious results because they greatly favour some sub-groups over others.

The use of common test approaches is also not restricted to comparability work in the UK. It is an approach that has been used quite widely in Australia, where subject attainment scores are routinely combined as part of university entrance procedures. In Queensland, for example, results from the Australian Scholastic Aptitude Test have been used as part of a procedure for scaling subject marks before combining them to produce tertiary entrance scores. In Sweden, comparability researchers have used the Swedish Scholastic Aptitude Test (SweSAT) within research studying grading standards over time in upper secondary schools (Wikstrom, 2005). Also, in the USA, year to year equating of the highly influential Scholastic Aptitude Test has depended to quite an extent upon the inclusion of a common element of 'anchor items' within successive versions of that national test. In none of these situations has the common test approach been seen as a perfect solution, but in each case it has been regarded as worth employing as one way of trying to get at elusive standard-setting goals.

We should note that statistical models cannot resolve the most fundamental dilemmas at the heart of the debates about the common test approach. The best that statistical treatments of this type can do is to indicate possible interpretations of findings, which need to be hedged with a variety of caveats concerning the need to treat the findings with caution and to remember that there are various interpretations that can be put on apparent discrepancies that may emerge from studies that attempt to investigate comparability questions using general ability measures as their benchmark.

As we shall see in subsequent sections there are fundamental issues that have arisen whenever this approach has been used in the last 40 years.

5 Evaluating the strengths and weaknesses of the common test approach

The previous section has both introduced some of the basic principles of common test comparability research, and inevitably it has also started to open up some of the difficulties that arise when such studies are subjected to rigorous scrutiny. We have also started to see that within what appears at first to be a fairly straightforward approach there are in fact many variations in how such studies can be designed, conducted and analysed. Furthermore, any judgement about the fitness for purpose of such studies of course depends crucially upon the questions that are being addressed by such studies.

As we shall see in a later section the early applications of the common test approach in the UK in the 1960s related to the introduction of the Certificate of Secondary Education (CSE) examinations, and a perceived need for the newly established CSE boards to be given some broad guidance, prior to grade awarding, about the relative 'ability levels' of the candidates entering for their exams (Willmott, 1980). That is a very different context from, for example, more recent standards over time debates,

where the common test approach has been used to try to indicate whether GCSE and A level grade awarding standards are being maintained from one year to the next (Tymms *et al*., 2005).

On the basis that we want to consider a balanced view of both the strengths and weaknesses of the common test approach, let us now summarise some of its advantages. Because of the nature of the challenge faced it is all too easy to dwell too much on the shortcomings and never properly acknowledge the advantages such an approach can have over others.

1. The common test approach to studying comparability is both easy to understand and represents an approach that on the face of it could be expected to yield worthwhile results. In a situation where different groups of students have sat different examinations and obtained different grades, knowing how they have all performed on a common test seems intuitively to be a worthwhile piece of information to be able to acquire.

2. If the common test is a relatively simple test to administer then the additional burden on the students and their schools is not colossal.

3. Common tests can sometimes be already available tests of general ability, so they don't have to be created specially and the same test can if necessary be used time and time again over a number of years. Indeed some common test studies (e.g. Massey *et al.*, 2003) have been able to identify existing test data, held in that case by Local Education Authorities, which could be called up and used without there being any need to retest the students.

4. Common test approaches, especially those based upon the use of general ability tests, can be applied to the more complicated comparability challenges such as changes in standards over long periods of time and the highly contentious area of between-subject comparability.

5. Finally, because all approaches to studying the comparability of examination grading standards have major limitations, the usefulness of the common test approach needs to be assessed in relation to the limitations of the other approaches. Hence it isn't just a matter of whether the common test approach is accurate enough, it is also a matter of whether, even with its shortcomings, it could provide better estimates than other flawed alternatives.

In subsequent sections we will continue to review other strengths and weaknesses of the common test approach. However, it is clear from overviewing the extensive literature that has been devoted to evaluating this method, that there are two key issues which arise most frequently. These relate firstly to the nature and relevance of the common test itself, and secondly to the extent to which a common test, when used, can be seen to offer a comparable challenge to specific sub-groups of students. We will now consider these two key issues in turn. First of all, we will consider the selection of the most appropriate type of common test for use in comparability studies.

Where common tests are employed to study the comparability of subject-based achievement tests such as GCSE and A level examinations, there is a real dilemma over whether to try to include subject content materials that will provide some direct overlap with the content of the examinations to be compared. The counter need is to avoid including specific test items that may bias the common test results towards the content of particular GCSE and A level syllabuses and examinations, and thus unbalance what is supposed to be a fair comparison. Although some early comparability researchers (Wrigley *et al*., 1967; Newbould & Shoesmith, 1974) did experiment with the creation of subject-based common tests, studies that looked at their performance alongside general ability tests tended to favour the latter rather than the former. It was found to be very hard to avoid syllabus-specific bias with such subject-based common tests. For this reason general ability tests have tended to be used in UK studies of comparability of examination grades since the mid-1970s.

The other factor that tends to favour general ability tests over subject-specific tests is the requirement in common test comparability studies to avoid bias in favour of sub-groups of candidates. Although general ability tests may not be wonderful predictors of subject-specific attainments, they are less likely than subject-specific tests to introduce bias through factors such as their coverage of syllabus areas. Nevertheless, there are other kinds of bias that can arise through using general ability tests as the comparator, and in some comparability studies it has been shown, for example, that they can provide very different predictions concerning the attainment of male and female candidates (Willmott, 1977).

We will look in more detail at some of the specific uses of general ability tests in comparability research studies in later sections. However, it is worth noting at this point that this characteristic of the common test approach is one of the key factors that tends to divide those who favour and those who criticise this type of research. The advocates of the general ability approach often argue that public examinations are in many respects all measures of general academic ability. Even though such examinations are each matched to closely defined subject-based syllabuses and assessment procedures, the highly specific nature of these assessments starts to be diluted as marks are combined from different assessment elements and turned into the grades that are awarded for the overall performance in that examination. They also go on to argue that most users of such examinations results are mostly interested in the way in which they give them a general overview of the students' academic abilities.

On the other hand those whom are less convinced by this approach tend to point to what they see as a serious inconsistency in trying to equate different subject-based examination results on the basis of estimates of the students' general ability. They generally go on to argue that where examinations are attempting to assess highly specific achievements, which reflect more about the quality of the students' learning and motivation in response to specific teaching approaches, it is a complete nonsense to assume that knowledge about their general academic ability will allow judgements to be made about the accuracy of their subject-based examination grades.

This debate about the appropriateness, or otherwise, of the use of general ability tests as the benchmark measure in common test studies is really quite hard to resolve. From a statistical point of view, general ability tests do generally correlate in a positive way with subject-based examination grades. However, such correlations are not terribly high and often fall in the range 0.3 to 0.7, indicating that although there is usually some similarity between general ability test scores and subject-based achievement test results, the relationship is not especially strong, and so it needs to be treated cautiously in the clear knowledge that many other factors are contributing to both sets of scores and the similarities and differences between them.

Even in a study where a general ability common test can be shown to be correlating quite well with the examination results (say of 0.7) there is still potential for plenty of debate about what such a statistical relationship implies. On the one hand, it is possible to argue that because both measures are themselves to some extent unreliable, then the differences between candidates' scores on the two measures can be explained largely in terms of random 'error' factors. On the other hand, it is possible to argue that such differences could reflect highly significant aspects of the subject-based achievements of the candidates, such as teaching, learning and motivation effects, which are highly meaningful and are in great danger of being distorted if any scaling were to be undertaken based upon general ability estimates.

Such a situation can be argued to be just the same as the situation where an examination has separate externally assessed and teacher assessed elements (Cohen & Deale, 1977). Some argue that teacher assessed elements should be scaled to reflect patterns observable in the externally assessed elements (Smith, 1978). Others have argued vehemently that to do such a thing totally undermines the credibility of teacher assessed judgements, which have their own validity and should not be subjected to correction factors that are derived entirely from comparisons being made with very different forms of student assessment (Macintosh, 1986).

Even the application of non-linear (Goldstein & Cresswell, 1996) and multilevel models (see Chapter 10) will not resolve this debate. Even if they provide improvements in the way in which data from common test studies can be analysed, they cannot ultimately resolve the dilemma over how discrepancies between general ability and subject-specific scores are interpreted and treated.

Here it is difficult to see any movement in the state of this debate from the point in time nearly 30 years ago when Bardell, Forrest & Shoesmith (1978) wrote:

> … to a large extent the… fairness of… [a common test] has to be subjective, since there is no way of distinguishing between biases in the monitor and the very differences in board standards which the exercise sets out to estimate.
>
> Bardell *et al.* (1978, p. 21)

6 An historical perspective on the use of the common test approach (1965 to 1985)

The period from 1965 to 1985 was undoubtedly the era during which the common test approach to comparability in England was most popular. This was also the period when the demand for comparability studies was arguably also at its highest point. There were during that period over 20 separate independent examining boards running CSE, O and A level and other public examinations. This high level of diversity of provision fuelled endless questions about the grades students were obtaining from different boards, subjects and types of examinations (Murphy & Torrance, 1988).

One major strand of comparability work at this time, emanated from the Examination and Test Research Unit (ETRU) at the National Foundation for Educational Research (NFER). That unit was established by the Schools Council to undertake work to support the establishment of the Certificate of Secondary Education (CSE) examinations, which were introduced in 1965. Comparability of assessment standards was the central preoccupation of the ETRU and the common test approach was used extensively by it in a series of investigations over a period of 12 years (Willmott, 1980). Many of the NFER studies used a general ability test as the point of comparison, although there was also some experimentation with subject-based common tests. In contrast, during the same period the GCE examining boards maintained a regular programme of inter-board comparability studies, which utilised a range of comparability research methods. Among the GCE board studies several used a common test approach, but unlike the NFER studies they often used subject attainment tests as the common element, and in quite a number of cases the subject attainment test was included as an integral part of the operational examinations (Bardell *et al*., 1978).

The very early work by the ETRU in the 1960s involved NFER researchers working closely with the newly established CSE boards to help them to decide how to create appropriate grading standards for the new CSE examination (Willmott, 1980). In fact the very first ETRU comparability study involved the use of a general scholastic aptitude test, which was taken by a sample of CSE candidates sufficiently early in the year, so that the ETRU could compare the common test results with teachers' predicted grades before CSE grades were finalised. On the basis of early analysis letters were sent to the CSE boards 'offering suggestions for standards of grading'. Then later on in the study the CSE grades actually awarded were analysed using the original common test results.

On reviewing the reports of the early ETRU studies in the late 1960s it is clear how there was a recognition of some of the shortcomings of the common test approach. Several attempts were made to improve the appropriateness of the common test and different aptitude tests were tried, sometimes alongside specifically constructed subject-based attainment tests (Wrigley *et al*., 1967; Skurnik & Connaughton, 1970). In general, any conclusions drawn from these early comparability studies were cautious and were presented alongside warnings about the assumptions upon which they

were based. Overall the main aim was to help the CSE boards to establish reasonably comparable grading standards.

Willmott (1980) recalls how the context for comparability research changed in the early 1970s. By this time the relaxed attitude towards such work in the late 1960s had changed to stronger public concerns about the grading standards being employed in CSE, O level and A level examinations. This move towards 'high stake comparability research' threatened the fragile basis upon which common test studies could be regarded as delivering accurate and dependable findings.

> The later studies, however, took place in times of growing public scrutiny of the public examination system and the reference test method of studying comparability simply could not stand up to – and indeed was not designed to stand up to – detailed scrutiny.
>
> Willmott (1980, p. 35)

This dramatic turn in the fortunes of the common test approach to comparability research led to the NFER and the Schools Council abandoning this whole strand of research and resulted in the ETRU being disbanded in 1977.

During this same period the GCE boards continued with a range of inter-board comparability studies utilising a mixture of cross-moderation and common test approaches. Their work differed from that of the NFER group in so far as it utilised a wide range of research approaches (Bardell *et al*., 1978). Another crucial difference between the GCE board studies and the NFER studies was that the board studies were conducted very much 'in-house' and were to inform those marking and grading the examinations being investigated. In contrast, the NFER studies were more public and led to quite a bit of press speculation about inadequacies in the standards of public examinations. Here we hit the problem of using an approach such as the common test approach to comparability in the knowledge that it has some major limitations that need to be taken into account in assessing any possible conclusions which might emerge from it. The simplicity of the approach can lead to some apparently simple conclusions about the comparability of examination grades – say between two different boards. However, a mass audience is more likely to latch on to the simple conclusions rather than the complicated caveats that might accompany them. It was this issue in particular that led to some fairly acrimonious exchanges between the NFER team and some examining board researchers.

> Dr. Willmott believes that it is better to produce results which may give rise to useful discussions of the assumptions, methodology and indeed the results themselves, than to consider that the problem is too difficult for study… You cannot do as Willmott has done and toss in results on a take-it-or-leave-it basis, much less over-interpret them. It is as if the Wright brothers had said, 'Our aeroplane doesn't fly yet but here is a pair of wings which you can use until it does'.
>
> Wood (1976)

As we have already noted the NFER team soon abandoned the common test approach after this exchange, and the GCE boards themselves started to rely much more on other methods for studying comparability from 1978 onwards (Forrest &

Shoesmith, 1985). It is particularly noticeable in the two published reviews of GCE board comparability work (Bardell *et al.*, 1978; Forrest & Shoesmith, 1985) how the emphasis moved strongly away from common test methods towards cross-moderation studies at this time.

7 The continued use of the common test approach (1986 to 2006)

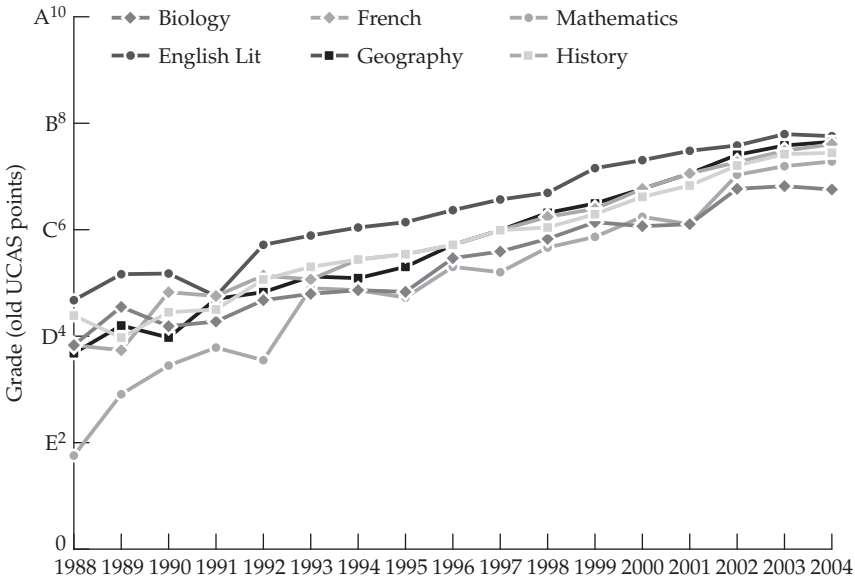
It seems that claims that the common test approach to comparability of examinations in England died in the 1970s 'have been greatly exaggerated'. Right up until the current time this approach has continued to be used especially by researchers from the Curriculum, Evaluation and Management Centre (CEM), which was located successively at the Universities of Newcastle and Durham. Although this group has been the most prominent user of this approach in recent years, it would be wrong to give the impression that they have been the only users of it. Occasional studies of comparability undertaken by University of Cambridge Local Examinations Syndicate (UCLES) researchers have continued to utilise it in situations where they judge it to have value (e.g. Dexter & Massey, 2000; Massey *et al.*, 2003). In addition others with an interest in debates about possible changes in examination standards over time, such as lecturers in higher education, have from time to time undertaken ad-hoc studies based upon requiring their students to take the same subject-based test over a number of years (Hunt & Lawson, 1996; Tariq, 2002).

Many of the same arguments apply to these studies, and the approaches used are in many respects similar to those used in the 1965–1985 period. Clearly most of the researchers involved are now more aware of the shortcomings of this approach, and their reports frequently contain relevant provisos and references to the limitations of their findings. For example, a study of A level standards by the Curriculum, Evaluation and Management (CEM) Centre, conducted for the GCE boards, acknowledged the following:

> …no single reference test relates in precisely the same way to all examinations and candidates from one examination may perform better on the reference test than candidates from another examination because working towards the examination is in some way a better preparation for the reference test. As a result scores on the reference test are not comparable for candidates entering the different examinations.
>
> GCE Examining Boards (1994)

In a more recent study Tymms *et al.* (2005) used the common test approach to look at evidence for changes over time during the period of office of the current Labour government. Figure 5 illustrates one aspect of their findings. Using data from their ALIS (A level information system) project they conducted various analyses to compare scores on a Test of Developed Abilities (TDA) with the A level grades obtained by the same students.

Figure 5 Mean UCAS grades achieved by students with a TDA score of 60



It is clear from Figure 5 that there has been a trend for students with the same score on the Test of Developed Ability (or the International Test of Developed Abilities, which they used before 2000) to achieve better A level results over this time period. Having discussed various possible explanations for this phenomenon, including the fact that it might be the result of better teaching, these authors conclude that:

> It is our view that A levels have generally become more leniently graded through a combination of syllabus change, modularisation and alterations to exam formats. In many ways that has been a good thing. It has allowed increasing numbers of candidates to access education to higher levels. But it has meant that the very top levels of attainment have been removed from A level.
>
> Tymms *et al*. (2005)

As with other common test studies the results from this study are amenable to all sorts of alternative interpretations. The authors themselves acknowledge the fact that causal explanations from the data represented in Figure 5 are far from straightforward. There are numerous possible factors that could have contributed to the trends, including syllabus changes, moves to modular assessments, improvements in teaching and pupil motivation and examination strategies, and indeed a modification of the general ability test itself, mid-way through the period. There is also the problem of their ALIS database containing samples of students who may not be representative, leading to the possibility that the trends reported here may not be reflected in the wider population of A level students.

So once again we have the common test approach yielding findings that on the face of it look straightforward, but that on closer inspection are far from simple to

interpret. On what basis should we expect A level grades to mirror students' scores on tests of general ability? How wise is it to discount teaching and learning effects and other factors such as motivation in studies of this kind? Finally, is enough attention being paid to the warnings of earlier researchers in relation to factors such as bias in the common test towards sub-groups of students taking the examinations being compared? There are few satisfactory answers to these questions, and any future use of the common test approach will always have to be seen as potentially defective unless such matters are properly addressed.

8 Towards a better informed use of the common test approach

In this chapter we have reviewed the use of the common test approach to studying the comparability of public examinations in England, Wales and Northern Ireland over a period of more than 40 years. Like other approaches to studying comparability this approach has both strengths and weaknesses.

Looking back over the common test studies that have been conducted during this time period, it is possible to conclude that this method has contributed in significant ways to the highly intractable challenges that comparability research has posed.

Common test studies seem to be most effective when they are conducted in relatively relaxed situations where results can be interpreted carefully and in the full knowledge of the various possible factors that can distort their findings. They can never produce simple answers to complex questions about the comparability of exams. However, they can certainly contribute useful evidence, which can be scrutinised alongside other findings from other methods in an attempt to piece together a case about the comparability of grading standards in alternative examinations.

The key element in the debate about the suitability of this method is undoubtedly the nature of the relationship between the common test and the examinations being compared. If those using the findings have a strong belief that the common test results should provide a firm indication of the results that should be expected from the examinations, then there is a greater likelihood that the results will be seen as significant. The only other challenge to validity then is the second key issue of the extent to which the common test results avoid favouring one sub-set of candidates over another.

There is no such thing as an easy comparability question. Even a question about the relative standing of grades awarded for two parallel forms of the same syllabus can take us into complex discussions about teaching effects. It is, however, possible to identify the most challenging comparability questions, such as the comparability of grades awarded for very different subjects, French and chemistry say, or the comparability of grades awarded for the same subject but 30 years apart. A strength of the common test approach is that it does provide a method for addressing even the most challenging of comparability questions. However, ascribing a sensible level of confidence to the findings of such a study is still a demanding prospect! So we can

use common general ability test methods to address both types of challenging comparability question, but the results may not move us very far towards answers in which we can have high levels of confidence.

9 Conclusion

In conclusion, it is possible to state the following as a summary of the key arguments covered in this chapter.

1. Common test methods are worth considering for inclusion in comparability studies, as long as their results can be treated with caution and due attention can be given to the various known threats to the validity of any apparent findings.

2. Those employing this approach must have a degree of confidence in the fact that the common test being used provides a worthwhile prediction of the results expected from the examinations to be compared.

3. Statistical checks need to be applied to the results of common test studies to ensure that the relationship between common test scores and examination results produces reasonably parallel regression lines.

4. Another condition that needs to be met is that the common test does not unduly favour any sub-group of candidates over others in such a way that between-examination comparisons are distorted.

5. In all studies of this kind it is necessary to indicate which definition of comparability is being used. This will help to clarify the relevance, say, of the use of a test of general ability, which may only be regarded as appropriate if comparability is assumed to involve students with the same level of general ability getting the same grades in subject-based public examinations.

6. All reports of future common tests comparability studies should address points 2–5 above, and include an explicit discussion of each issue in order that any users of the findings are fully aware of all major threats to the validity of any conclusions drawn.

A broad conclusion to this review is that things have changed very little since Newbould & Massey (1979) carried out their comprehensive review of this approach from which they concluded that 'Common tests would seem to look more attractive monitors of standards than they really are' (Newbould & Massey, 1979, p. 51).

None of what has happened in the years since that report was written has changed the appropriateness of their overall conclusion that:

> Perhaps the common test, from this particular standpoint, is on balance, harmful, for whilst it is a ready vehicle for emphasising common ground, and hence implying a similarity of examinations, it may tempt people into adopting over-simplistic views of the nature and meaning of the concept of comparability.
>
> Newbould & Massey (1979, p. 51)

Those administering and researching public examinations need to be highly aware of the impact of their actions on public perceptions (Warmington & Murphy, 2004; 2007) and common test approaches to comparability can be seen to be at risk of encouraging oversimplistic views about comparability. Our very sophisticated system of public examinations demands a similarly sophisticated basis for judging the dependability and meaningfulness of candidate results. An approach based upon the common test method isn't likely to ever be described as sophisticated. Comparability studies involve highly complex comparisons between areas of educational achievement, which have little in common with each other. Such comparisons need us to cope with levels of uncertainty found in expert judgements (Murphy, 2004), and an acceptance that many aspects of educational achievement transcend simplistic comparisons on a uni-dimensional scale.

References

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., Fearnley, A., Fowles, D., Jones, B., Morfidi, E., & While, D. (2001). *Tiering in the GCSE*. London: Joint Council for General Qualifications.

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.

Black, P. (1998). *Testing: Friend or foe? Theory and practice of assessment and testing*. London: Falmer Press.

Broadfoot, P.M. (1996). *Education, assessment and society: A sociological analysis*. Buckingham: Open University Press.

Brownlee, K. (1965). *Statistical theory and methodology in science and engineering*. New York: Wiley.

Cohen, L., & Deale, R.N. (1977). *Assessment by teachers in examinations at 16+*. Schools Council Examinations Bulletin 37. London: Evans/Methuen

Dexter, T., & Massey, A., (2000, July). *Conceptual issues arising from a comparability study relating IGCSE grading standards with those of GCSE via a reference test using a multilevel model*. Paper presented at the 22nd biennial conference of the Society for Multivariate Analysis in the Behavioural Sciences at the London School of Economics, London.

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.

GCE Examining Boards. (1994). *Comparing examination boards and syllabuses at A level:*

Students' grades, attitudes and perceptions of classroom processes. Executive summary of a report commissioned by the GCE Examining Boards. Belfast: Northern Ireland Council for the Curriculum, Examinations and Assessment.

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

Goldstein, H., & Cresswell, M.J. (1996). The comparability of different subjects in public examinations: A theoretical and practical critique. *Oxford Review of Education, 22*, 435–441.

Hunt, D.N., & Lawson, D.A. (1996). Trends in mathematical competence of A level students on entry to university. *Teaching Mathematics and its Applications*, *15*(4), 167–173.

Macintosh, H. (1986). The sacred cows of coursework. In C. Gipps (Ed.), *The GCSE: An uncommon examination*. Bedford Way Papers 29. London: University of London Institute of Education.

Massey, A., Green, S., Dexter, T., & Hamnett, L. (2003). *Comparability of national tests over time: Key stage test standards between 1996 and 2001.* London: Qualifications and Curriculum Authority.

Meadows, M. (2003). *Comparability of the biology and human biology routes to GCSE biology A 2002*. Unpublished Research Report, Assessment and Qualifications Alliance.

Murphy, R.J.L. (2004). *Grades of uncertainty*. London: Association of Teachers and Lecturers.

Murphy, R.J.L., & Torrance, H. (1988). *The changing face of educational assessment*. Milton Keynes: Open University Press.

Murphy, R.J.L., Wilmut, J., & Wood, R. (1996). Monitoring A level standards: Tests, grades and other approximations. *The Curriculum Journal*, *7*, 279–291.

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

Newbould, C.A., & Shoesmith, D.J. (1974). *Technical drawing reference test 1973*. Appendix A to the Report on the 16+ Feasibility Study in Technical Drawing by the East Anglian Examination Board and University of Cambridge Local Examinations Syndicate. Cambridge: University of Cambridge Local Examinations Syndicate.

Newton, P.E. (1997). Examining standards over time. *Research Papers in Education*, *12*, 227–248.

Nuttall, D.L. (1979). The myth of comparability. *Journal of National Association of Inspectors and Advisors, 11*, 16–18.

Skurnik, L.S., & Connaughton, I.M. (1970). *The 1967 CSE monitoring experiment*. Schools Council Working Paper 30. Evans/Methuen Education.

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

Smith, G.A. (1978). *JMB Experience of the moderation of internal assessments*. Occasional Publication 38. Manchester: Joint Matriculation Board.

Stringer, N. (2005). *Response to scrutiny report for GCSE PE and PE games*. Unpublished Research Report, Assessment and Qualifications Alliance.

Tariq, V.N. (2002). A decline in numeracy skills among bioscience undergraduates. *Journal of Biological Education*, *36*(2), 76–83.

Tymms, P., Coe, R., & Merrell, C. (2005, April). *Standards in English schools: Changes since 1997 and the impact of government policies and initiatives*. A report for *The Sunday Times*, London.

Warmington, P., & Murphy, R. (2004). Could do better? Media depictions of UK assessment results. *Journal of Education Policy, 19*, 285–300.

Warmington, P., & Murphy, R. (2007). Read all about it! UK news media coverage of A level results. *Policy Futures in Education*, *5*(1), 70–83.

Wikstrom, C. (2005). Grade stability in a criterion-referenced grading system: The Swedish example. *Assessment in Education*, *12*, 25–144.

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

Wilmott, A.S. (1980). *Twelve years of examinations research: ETRU 1965–1977*. London: Schools Council.

Wood, R. (1976, July 30). Your chemistry equals my French. *The Times Educational Supplement*.

Wrigley, J., Sparrow, F.H., & Inglis, F.L. (1967). *Standards in CSE and GCE (English and mathematics)*. Schools Council Working Paper 9. London: Her Majesty's Stationery Office.

COMMENTARY ON CHAPTER 8

Robert Coe, Peter Tymms and Carol Fitz-Gibbon

In his chapter on common test methods, Roger Murphy presents a critique of the approach. Though he does acknowledge strengths as well as weaknesses of the method, his view seems to be that the latter are more pertinent. Our view is that the weaknesses of this method have been exaggerated and some of its potential strengths overlooked. Furthermore, a few of the criticisms presented in the chapter are simply unjustified. In this commentary we attempt to redress this imbalance and defend the use of common test methods. There are six specific issues on which we comment.

The worst form of comparability research?

Winston Churchill said that 'democracy is the worst form of government, except all those other forms that have been tried from time to time'.[1] As in government, so in comparability research; an approach that is easy to criticise may nevertheless be the best we can do. Like any method applied to the complex business of trying to compare standards across different examinations, common test methods have their limitations. If we use them we must be aware of these limitations, be sensitive to the kinds of assumptions required for their application and be cautious about any claims we make.

But whatever the problems of the common test approach may be, the alternatives to using these methods (including not asking any questions about comparability in the first place) all have their problems too. Murphy seems to acknowledge this, stating that 'all approaches to studying the comparability of grading standards have major limitations', and that 'it is all too easy to dwell on the shortcomings'. However, it is hard to resist the impression that his view of common test methods is, on balance, negative. His position is indicated by describing the approach as 'potentially defective' and confirmed by a concluding quotation that it is, 'on balance, harmful'. What is less clear is whether common test approaches are worse than other statistical methods, whether all statistical approaches, including common tests, are inferior to judgement methods, or even whether all attempts to address the comparability of different examinations by whatever method are 'on balance, harmful'.

In fact, none of these positions is helpful. We cannot talk about the limitations of a method per se, but about the kinds of claims, interpretations and uses it can validly support. On its own a method is neither valid nor invalid; it depends how its results are interpreted. Murphy seems to acknowledge this, accusing common test methods of 'encouraging oversimplistic views' and advising that they are 'worth considering… as long as their results can be treated with caution'. However, he fails to give any examples of uses of the method that he believes to be appropriate.

There are different conceptions of comparability

The issue of the different meanings of the word 'comparability' is treated rather briefly in the chapter, but is a crucial one for understanding the appropriateness of using common test methods. This matter is explored in the commentary on Chapter 4.

Murphy's specification of the requirements for the use of common tests that there should be 'reasonably parallel regression lines' and that the test should not 'unduly favour any sub-group' suggests that he sees their use as limited to comparing standards on examinations that are essentially equivalent forms, measuring the same construct in the same ways. These conditions might be met, for example, in comparing two examinations in the same subject from different boards, but may or may not be true for two examinations in different subjects.

However, other parts of the chapter seem to acknowledge that a common test might also be used within a 'value-added' conception of comparability, as, for example, in the assumptions underpinning conclusion 5. In such a view of comparability, if the common test were a measure of prior attainment or ability, a regression model could be used to estimate the gains made by similar students in different examinations. If students consistently achieve more in one examination from the same starting point, we might conclude, *ceteris paribus*, that it is more leniently graded, i.e. easier. Comparability here is defined in terms of equal gains. This conception of comparability does not depend on parallel lines or subgroup invariance, though it does require us to define what we mean by 'similar students' and to make the additional assumption that other factors (quality of teaching, effort applied, etc.) are either equal or irrelevant. The point here is not that these alternative assumptions are less problematic – they may well not be – but simply that they are different.

Yet another conception of comparability would see the common test as the linking construct in terms of which a set of examinations may be compared. In this case, the requirements are different again. An example of this conception of comparability may be found in the study by Tymms, *et al.* (2005), referred to by Murphy. His suggestion, that factors such as 'syllabus changes, moves to modular assessments, improvements in teaching and pupil motivation and examination strategies' might account for the changes, may have some validity, but does not alter the interpretation in this case. In a 'construct comparability' conception, examinations in different years are compared on the basis of the level of the linking construct, general ability, to which particular grades correspond. Whatever the reason for increasingly higher grades to be awarded to candidates of the same ability – including any of those mentioned above – it remains the case that the award of the same grade denotes progressively lower abilities (as measured by the common test) each year.

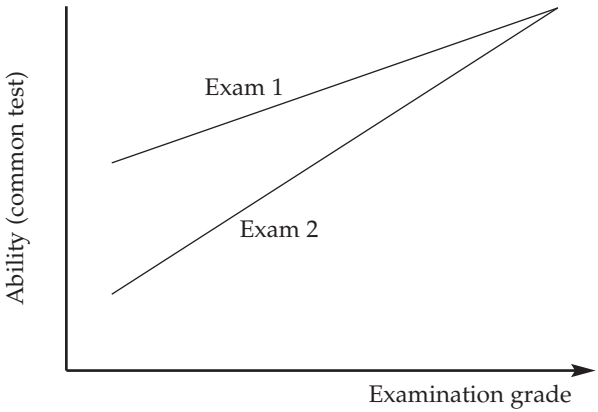
Parallelism may not be a requirement

Murphy argues that if the regression lines for two examinations are not reasonably parallel then that 'poses a serious threat to the validity of any conclusions drawn from a study of this kind.' This is because 'the common test may have different

relevance for the candidates in the two groups being compared.' In his Figure 2, the gradients of the two regression lines shown are so different as to be a caricature of the kinds of graphs that are routinely found. To suggest that such a finding is 'far from uncommon' is misleading since such an excessive difference would indeed be extremely uncommon.

However, if a common test is interpreted as a linking construct, even such a difference might not matter at all. Consider the hypothetical graph shown in Figure 1.

Figure 1



Here the regression lines are clearly not parallel. This kind of situation would arise, for example, if candidates for Exam 1 had a more limited range of ability than those for Exam 2, but the same range of grades was awarded for each. However, whether this means that the common test has 'different relevance' for the two may be questioned; it may simply be a case of some stretching of the scale on which examinations are reported. With a 'linking construct' interpretation of comparability, the question we might ask about the two examinations is: 'What level of ability corresponds to the award of a particular grade in each examination?' The fact that the lines are not parallel is no barrier to asking – or answering – this question. We therefore believe that Murphy's conclusion 3 is unsound.

Statistical significance and the size of the difference

Murphy gives an example from 1969 of a comparison between CSE English examinations from different boards which failed to show a statistically significant difference between them. He goes on to say that such studies 'will never give highly accurate outcomes', and that it is 'foolish to expect high levels of precision from research studies which seek to compare different examinations.'

In fact, of course, given a large enough sample, one can achieve any level of accuracy desired. The fact that a mean-on-mean comparison with n=9 failed to reach the traditional (but quite arbitrary) level of statistical significance (by this method)

conceals the fact that there appear to be some very substantial differences between these boards. Interestingly, the converse of this error is made by Newbould & Massey (1979) who apply the common test method to comparisons of different boards. They draw a series of regression lines that to the eye seem very close to parallel, but, because they have a sufficiently large sample, are shown to be statistically 'significantly' different. A clear distinction between statistical and educational significance is important.

How strong a correlation is required?

Murphy suggests that correlations below about 0.7 are too low for the common test method to be valid. It is not clear, however, what the justification is for the choice of this particular value. Correlations of the order of 0.7 are common in value-added analyses, but this does not seem a very good reason for judging them acceptable.

One issue that does not appear to have been considered is the effect of any range restrictions. For example, some of the correlations cited by Coe (1999) do indeed appear quite low. However, if one corrects them for the restricted range of candidates taking these A level examinations, a measured correlation of 0.4 turns out to be equivalent to about 0.7 in the full population, in terms of the strength of the relationship between the two variables.

A second issue relates to the reliability of the examinations – lower reliabilities produce lower correlations and since no tests are perfect all the reported correlations underestimate the underlying relationships.

But neither of these issues helps in answering the question about what is an acceptable correlation. A starting point would have to be a statement as to what type of comparability was being addressed. Then a power calculation could follow which took off from a further statement of the kinds of errors which would be acceptable in any proposed study. To our knowledge this kind of calculation has never been carried out.

Publication is democratic

Murphy makes a number of references in the chapter to the problems of making research on comparability available to a 'mass audience'. He argues that caveats, subtleties and 'our very sophisticated system of public examinations' are hard to convey in public reports, and these studies 'seem to be most effective when they are conducted in relatively relaxed situations', presumably away from the kind of public controversy that often seems to accompany publication of such studies.

We would certainly agree that researchers should make the limitations of their work clear and not oversimplify, giving due attention to threats to validity and alternative explanations. However, it would be wrong to interpret this as meaning that research on comparability should not be made available to a mass audience. The fact that the media may sometimes misrepresent findings is not a reason to confine presentation of them to internal awarding body reports or to seminars attended by only the

enlightened few. Such secrecy would be against the academic ideal of open critical debate and the democratic principle of transparency in public institutions. It would also deny those researchers what in our experience have been some of the most valuable insights into these issues that arise from a more public debate.

Endnote

1 Speech in the House of Commons, 11 November 1947.

References

Coe, R. (1999, September). *Changes in examination grades over time: Is the same worth less?* Paper presented at the British Educational Research Association annual conference, Brighton.

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

Tymms, P., Coe, R., & Merrell, C. (2005, April). *Standards in English schools: Changes since 1997 and the impact of government policies and initiatives*. A report for *The Sunday Times*, London.

RESPONSE TO COMMENTARY ON CHAPTER 8

Roger Murphy

The commentary on this chapter comes from a group of individuals who are recognisable as enthusiastic users of the common test approach. In my view they have both misrepresented and misunderstood the very serious critique of this method, which I have presented. I did not attempt to compare all comparability approaches, but if I had I certainly would not have agreed with their implied view that the common test approach is 'the best we can do'.

In the commentary I am portrayed as not wanting comparability questions to be addressed (wrong), not wanting research findings to be made publicly available (wrong), and being dogmatic about the statistical properties required of common test studies (again wrong – where for example was I supposed to have said that common test with examination correlations must be higher than 0.7!). My stance in this chapter, and elsewhere, continues to be that assessment data, including statistical assessment research studies, can have serious limitations and that these need to be closely attended to, when drawing conclusions, whether about the findings of comparability research studies or even about the examination results of a single student (Murphy, 2004).

These commentators appear wedded to a view of comparability based entirely upon regarding the 'ability' of students as being the most important deciding factor when it comes to comparing their performance on different examinations. Coe (1999) illustrates this view, when he states that 'it matters less whether the award of each grade represents the same achievement than whether it signifies the same level of general ability in the candidate'. The implication here is that if two examinations give equal grades to students, who demonstrate similar levels of achievement, then those grades should be seen as inappropriate if the levels of general ability of the students getting them are different. Thus one difference between us is that I view public examinations as being measures of educational achievement in relation to highly specific areas of the curriculum, whereas they see such grades as some kind of proxy for students' levels of general ability. This is a very limiting stance to adopt because it does not leave scope for students to achieve more as a result of good teaching or their own hard work – the very essence of what educators are striving for!

For this reason I do continue to stand by the list of issues which I think are essential for future users of the common test approach. Comparability research is about so much more than obsessively checking the general ability levels of students in order to compare examination grades. This commentary depends fundamentally on a 'linking concept', which is in effect no more than the old Spearman/Binet/Burt belief in 'basic intelligence' being the fundamental principle around which all education

systems should be arranged. Thankfully educational practices in the 21st century have moved well away from that hugely limiting view of the world, and comparability research has much more to offer than that view has ever provided or ever will provide.

References

Coe, R. (1999, September). *Changes in examination grades over time: Is the same worth less?* Paper presented at the British Educational Research Association annual conference, Brighton.

Murphy, R.J.L. (2004). *Grades of uncertainty*. London: Association of Teachers and Lecturers.