Techniques for monitoring the comparability of examination standards, edited by: Paul Newton, Jo-Anne Baird, Harvey Goldstein, Helen Patrick and Peter Tymms

7

PAIRED COMPARISON METHODS

Tom Bramley

Abstract

Aim

The aims of this chapter are:

1. to explain the theoretical basis of the paired comparison method
2. to describe how it has been used in the cross-moderation strand of inter-board comparability studies
3. to discuss its strengths and weaknesses as a judgemental method of assessing comparability.

Definition of comparability

The chapter follows the approach of Hambleton (2001) in distinguishing between content standards and performance standards. Cross-moderation exercises are shown to be comparisons of performance standards between the examining boards. The judges are expected to make judgements about relative performance standards in the context of possible differences in content standards. In this chapter the performance standard is conceived as a point on a psychometric latent trait.

Comparability methods

In a paired comparison task a judge is presented with a pair of objects and asked to state which possesses more of a specified attribute. In the context of this chapter, the objects are examinees' scripts on a specified grade boundary from different examining boards, and the attribute is 'quality of performance'. Repeated comparisons between different pairs of scripts across judges allows the construction of a psychological scale (trait) of 'perceived quality'. Each script's location on this scale depends both on the proportion of times it won and lost its comparisons, and on the scale location of the scripts it was compared with. Differences in the mean location of scripts from the different boards have been taken to imply a lack of comparability – that is, differences in performance standards.

The chapter also describes a modification of the method to use rank-ordering rather than paired comparisons to collect the judgements (Bramley, 2005a). The underlying theory and method of analysis are the same.

History of use

The psychometric theory underlying the paired comparison method was developed by the American psychologist Louis Thurstone, who used it to investigate a wide range of psychological attributes (e.g. 'seriousness of crime'). It was first used in comparability studies to compare some of the (then) new modular A level syllabuses against their linear equivalents (D'Arcy, 1997), and since then has been the favoured method for the cross-moderation strand of inter-board comparability studies, which are the focus of this chapter. It has also been used to investigate comparability of standards over time, and, in its rank-ordering guise, as a technique for standard maintaining – enabling a known cut-score on one test to be mapped to an equivalent cut-score on a new test.

Strengths and weaknesses

The method has several advantages in the cross-moderation context. First, the individual severities of the judges are experimentally removed – that is, it does not matter how good (in absolute terms) they think the scripts they are judging are: all that matters is their relative merit. Second, the analysis model naturally handles missing data because the estimate of the scale separation between any two scripts does not depend on which other scripts they are compared with. This means that data can be missing in a non-random way without affecting the results. Third, fitting an explicit model (the Rasch model) to the data allows investigation of residuals to detect misfitting scripts and judges, and judge bias. Finally, the approach is simple and flexible, allowing the design of the study to be tailored to the needs of the particular situation.

One drawback to using the method in this context is its psychological validity when the objects to be judged are as complex as scripts. In Thurstone's own work the judgements could be made immediately, but here a certain amount of reading time is required. Also, the method assumes that each comparison is independent of the others, but this seems implausible given that judges are likely to remember particular scripts when they encounter them in subsequent comparisons. Unfortunately the paired comparison method is tedious and time-consuming for the judges, a drawback that can be remedied to some extent by using the rank-ordering method.

Conclusion

The paired comparison method of constructing psychological scales based on human judgements is well established in psychometric theory and has many attractive features which have led to its adoption as the preferred method in inter-board comparability studies. Many of the issues arising with the method are due to this particular context for its application. In practice, the most serious problem has not been with the method but with the design of the studies, which have not allowed differences between boards in terms of mean scale location to be related to the raw mark scales of the different examinations. This has made it impossible to assess the importance of any differences discovered (for example in terms of implied changes to grade boundaries). Both the paired comparison method and especially the rank-ordering method could easily address this shortcoming in future studies.

1 Introduction

The method of paired comparisons has been used in the cross-moderation strand of inter-board comparability studies for around a decade. The first studies to use it were reported in D'Arcy (1997). It replaced the cross-moderation techniques of *ratification* (where judges decide whether a script is above, at, or below a particular grade boundary) and *identification* (where judges examine a set of scripts in a range around the presumed boundary and identify the mark that best represents the grade boundary). These techniques, and the studies that used them, have been described in detail in Chapter 6. The cross-moderation strand is the part of a comparability study that requires expert judgements about examinee performance. The use of paired comparisons has not diminished the role of the specification review strand, which requires expert judgements about the level of demand of syllabuses and question papers (see Chapter 5) – this work is as important a pre-requisite for paired comparison judgements as it is for ratification or identification.

Section 2 of this chapter reviews some of the work of the American psychologist L. L. Thurstone (1887–1955), who established and developed the theoretical approach to psychological measurement that underpins the use of paired comparisons and related methods. Section 3 shows how the paired comparison method has been used in practice in comparability research in British examinations. Section 4 describes a rank-ordering method – a recent development based on the same underlying theory and using similar techniques of analysis. Finally, section 5 contains more general discussion about the nature and validity of cross-moderation exercises.

2 Background

The method of paired comparisons is a simple and direct way of collecting judgement data. The judge is presented with two objects (or 'stimuli') and has to decide which object is 'x-er' – in other words which object possesses more of a specified attribute, 'x'. In early psychophysical research 'x' was often a physical attribute such as weight, loudness or brightness. One aim of such research might be to discover or verify a mathematical function linking the perceived magnitude of the objects to their actual measured physical magnitude. A second aim might be to discover the 'limen' – the smallest increment in stimulus magnitude that could be discriminated at an arbitrary level of accuracy.

Two well-known results from this field of research are Fechner's law and Weber's law. Fechner's law states that the perceived magnitude of a stimulus is proportional to the logarithm of its physical magnitude, and Weber's law states that the smallest perceptible difference in stimulus magnitude is proportional to the absolute magnitude. Thurstone liked to quote the psychologist William James's view of psychophysics:

> William James said that psychophysics was the dullest part of psychology. He was right. But if we extend and adapt the psychophysical concepts to the theory of discriminatory and selective judgment, the subject takes a different colour and it is no longer dull.
>
> Thurstone (1945)

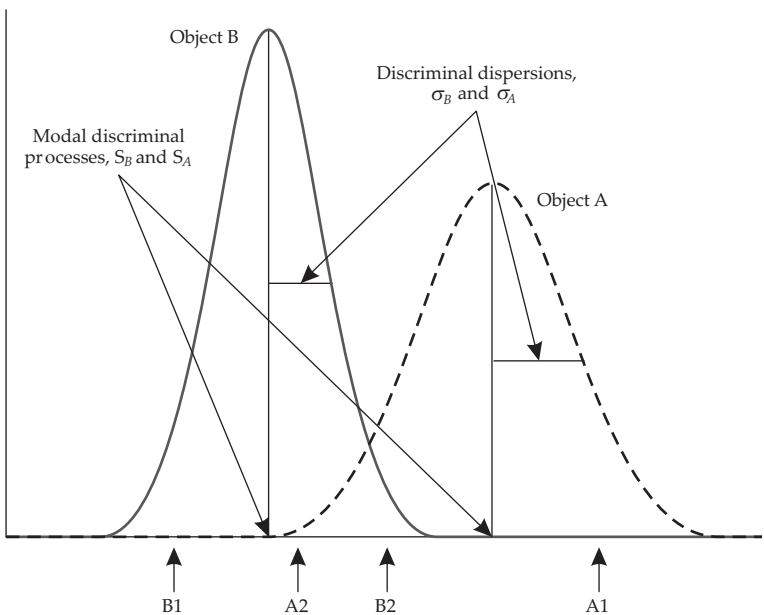
One of Thurstone's main achievements was to liberate psychophysical judgements from being restricted to attributes of stimuli that had a measurable physical magnitude and to develop a theory of psychological or 'subjective' measurement for non-physical attributes such as 'seriousness of crimes', 'attitude towards gambling', 'excellence of handwriting' and so on. This theory was first expounded in a series of articles published between 1925 and 1935, and Thurstone was still refining his theory in the years shortly before his death in 1955.

The key elements in Thurstone's theory are the 'discriminal dispersion' and the 'law of comparative judgement'. These are described below.

2.1 The discriminal dispersion

Thurstone assumed that each time a judge encounters a stimulus, it has some kind of psychological impact. He called this impact the 'discriminal process' – a term designed to be as neutral as possible regarding what is actually happening in the judge's mind or brain. He further assumed that the same stimulus is not necessarily associated with the same discriminal process each time the subject encounters it, but that there is a 'discriminal dispersion' or distribution of frequencies with which a given stimulus is associated with a particular discriminal process. The mode of this frequency distribution (the 'modal discriminal process') defines the location of the stimulus on the psychological continuum, and the standard deviation defines an arbitrary unit of measurement. The measurement scale is constructed on the basis that the distribution of discriminal processes is Normal (Gaussian).

Figure 1 Example distributions of discriminal processes for objects A and B



Different stimuli will have different modal discriminal processes on the psychological continuum and the distance between these modes corresponds to the scale separation of the stimuli. Two hypothetical distributions of discriminal processes are shown in Figure 1. The modal discriminal processes for stimuli A and B are located at points S_A and S_B on the psychological scale, with discriminal dispersions σ_A and σ_B respectively. It is crucial to the construction and definition of the psychological continuum that there is overlap between the distributions of discriminal processes along the continuum – in other words that not only can the same stimulus give rise to different discriminal processes, but that the same discriminal process can be evoked by different stimuli.

Thurstone was under no illusions about the nature of the measurement scales his methods produced. He clearly recognised that they existed in their own right in the psychological domain and was at pains to point out that his theory was neutral about whether psychological processes could be related to physiological processes, as the following two extended quotations show.

> The psychological continuum or scale is so constructed or defined that the frequencies of the respective discriminal processes for any given stimulus form a normal distribution on the psychological scale. This involves no assumption of a normal distribution or of anything else. The psychological scale is at best an artificial construct. If it has any physical reality, we certainly have not the remotest idea of what it may be like. We do not assume, therefore, that the distribution of discriminal processes is normal on the scale because that would imply the scale is there already. We define the scale in terms of the frequencies of the discriminal processes for any stimulus. This artificial construct, the psychological scale, is so spaced off that the frequencies of the discriminal processes for any given stimulus form a normal distribution on the scale.
>
> Thurstone (1927b)

> Any perceptual quality which may be allocated to a point on the psychological continuum is not itself a magnitude. It is not divisible into parts. It is not a sum of any mental or physical units. It is not twice, three times, or four times as strong, high, beautiful, or good as some other process on the same continuum. It is not a number. It is not a quantity… With these negations granted, just how do these qualitative entities or processes become a measurable continuum? They acquire conceptual linearity and measurability in the probability with which each one of them may be expected to associate with any prescribed stimulus.
>
> Thurstone (1927c)

Thurstone regarded the concept of the discriminal dispersion as an important theoretical innovation. For example, he used it to show that equally often noticed differences did not necessarily correspond to equal differences on the psychological scale (Thurstone, 1927e); that Fechner's law and Weber's law could be described in the same algebraic framework as his own law of comparative judgement (Thurstone, 1927f); and that it could explain some problems in the prediction of choice, such as voting behaviour (Thurstone, 1945). It is clear that he did not expect the discriminal dispersions of all objects to be the same if the attribute being judged was complex:

> It is probably true that this variability of the discriminal dispersions on the psychological

> continuum is of relatively less serious importance in dealing with strictly homogenous stimulus series, but it becomes a serious factor in dealing with less conspicuous attributes or with less homogenous stimulus series such as handwriting specimens, English compositions, sewing samples, oriental rugs.
>
> Thurstone (1927a)

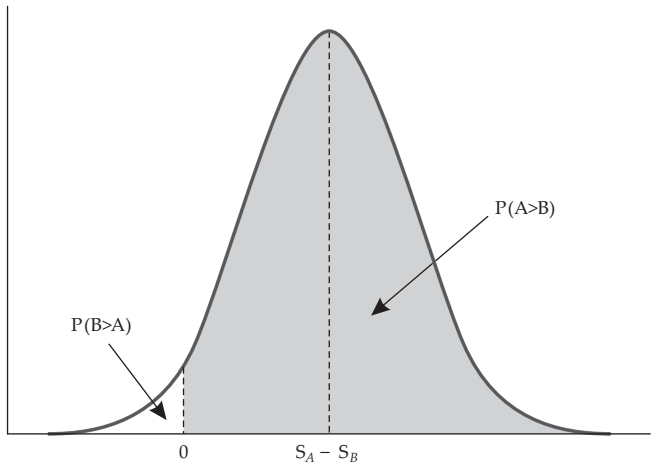
This issue of variability of discriminal dispersions will be considered later.

2.2 The law of comparative judgement

Thurstone linked his psychological theory to experimental data through the law of comparative judgement (Thurstone, 1927b). The discriminal mode and dispersion corresponding to a single stimulus are inaccessible to observation. They can only be estimated when two objects are compared. Thurstone assumed that when two objects are compared with respect to a specified attribute, the object evoking the discriminal process further along the psychological continuum would be judged as possessing more of the attribute. For simplicity assume that the attribute is 'quality of handwriting' and the objects are student essays. For example, in Figure 1, if essay A evoked the discriminal process at A1 and essay B evoked the discriminal process at B1 then essay A would be judged as having better handwriting than essay B. In contrast, if essay A evoked the discriminal process at A2 and essay B evoked the discriminal process at B2 then essay B would be judged as having better handwriting than essay A. It is clear from Figure 1 that the proportion of judgements 'A better than B' is likely to be much higher than the proportion 'B better than A' because the latter can only happen in the small range of overlap between the two distributions of discriminal processes.

The outcome of the paired comparison judgement is therefore related to the distribution of the *difference* between the two distributions of discriminal processes for essay A and essay B. If this difference is positive, we have the judgement 'A beats B', and if it is negative we have the judgement 'B beats A'. The distribution of differences is shown in Figure 2 below.

Figure 2 Distribution of discriminal differences between two objects, A and B



The mean of this distribution is the distance between the two mean discriminal processes – that is, the scale separation between A and B. The standard deviation of this distribution, σ_{AB}, is given by the formula:

$$\sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2 - 2 \cdot r_{AB} \cdot \sigma_A \cdot \sigma_B} \qquad (1)$$

where:

σ_A is the discriminal dispersion for essay A

σ_B is the discriminal dispersion for essay B

r_{AB} is the correlation between the discriminal processes.

The shaded area of Figure 2 to the right of the zero thus corresponds to the proportion of judgements 'A better than B'. This gives the law of comparative judgement:

$$X_{AB} = \frac{S_A - S_B}{\sigma_{AB}} \qquad (2)$$

where X_{AB} is the deviate of the Normal distribution corresponding to the proportion of judgements 'A beats B', and S_A, S_B and σ_{AB} are as defined above.

In words, the scale separation between two objects on the psychological continuum is measured in units of the standard deviation of the difference between the distributions of their discriminal processes.

Equation (2) is the most general form of Thurstone's law of comparative judgement. It applies to repeated comparisons by the same judge of the same pair of objects. However, in order to estimate the unknown quantities in the equation various simplifying assumptions need to be made. Thurstone identified five 'cases' of his law, each requiring more assumptions than the last.

Case 1 merely requires the assumption that the correlation, *r*, between pairs of discriminal processes is constant for all pairs of objects. Without this assumption no parameters can be estimated because each pair of objects introduces a new 'unknown' into the equation.

Case 2 makes the much bigger assumption that the same equation can be applied to a group situation – in other words instead of the proportion of 'A beats B' coming from replications within the same judge, it comes from replications across judges. Whereas within an individual the discriminal processes are Normally distributed on the psychological scale by definition, this Normal distribution becomes an assumption

when applied to the distribution of a single discriminal process in each judge across a group of judges.

Case 3 simplifies by assuming that the correlation term *r* is zero. Thurstone justified this simplification by identifying two opposing factors at work in a paired comparison – 'mood' and 'simultaneous contrast' – which might cancel each other out. The 'mood' factor would be exemplified by both objects in a pair evoking discriminal processes above the mode when a judge is in a 'generous' mood, and below the mode when the judge is in a 'mean' mood, giving rise to a positive correlation and hence a positive non-zero value for *r*. 'Simultaneous contrast', on the other hand, occurs when the difference between objects is perceived in an exaggerated way (for example, a tall person might appear taller and a short person appear shorter when they are standing next to each other than when standing on their own). This would give rise to a negative correlation between the discriminal processes, and hence a negative non-zero value for *r*, counteracting the mood effect. However, Andrich (1978) showed that if a 'judge effect' across judges is parameterised, it is eliminated experimentally in the paired comparison method (see section 3). Hence, for Case 3, the denominator of equation (2) simplifies to:

$$\sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2} \qquad (3)$$

Case 4 simplifies further by assuming that the discriminal dispersions are all fairly similar, which allows the denominator of equation (2) to be simplified to:

$$\sigma_{AB} = \frac{(\sigma_A + \sigma_B)}{\sqrt{2}} \qquad (4)$$

Case 5 makes the greatest simplification by assuming that the discriminal dispersions are all equal, which further simplifies the denominator of equation (2) to:

$$\sigma_{AB} = \sqrt{2} \cdot \sigma \qquad (5)$$

which, if the constant denominator σ_{AB} is treated as the (arbitrary) unit of measurement, means that:

$$X_{AB} = S_A - S_B \qquad (6)$$

In words: the scale separation between two objects is equal to the unit Normal deviate corresponding to the proportion of judgements 'A better than B'. (Note that if this proportion is less than 0.5 the separation will be negative – that is, B will be higher up the scale than A, as we would expect.)

Thurstone seems to have had a rather ambivalent attitude to the Case 5 version of his law, saying 'This is a simple observation equation which may be used for rather

coarse scaling' (Thurstone, 1927b), yet it seems to have been the one he used most often in practical applications!

It will perhaps not come as a great surprise to the reader to discover that it is the Case 5 version of Thurstone's law that has usually been used in cross-moderation exercises. The only difference is in the mathematical function linking the observed proportions to the difference in scale values – the more tractable logistic function is used rather than the cumulative Normal.

Equation (2) can be rewritten as:

$$p(A > B) = \frac{1}{\sigma_{AB}\sqrt{2\pi}} \int_0^{\infty} \exp\left(-\frac{[t-(S_A - S_B)]^2}{2\sigma_{AB}^2}\right) dt \qquad (7)$$

where *p(A>B)* is the probability that object A beats object B in a paired comparison, and *t* is the scale separation of the discriminal processes evoked by A and B in a single comparison.

The logistic equivalent is:

$$p(A > B) = \frac{\exp[a(S_A - S_B)]}{1 + \exp[a(S_A - S_B)]} \qquad (8)$$

where *a* is a scaling parameter, which can arbitrarily be set to 1 (just as σ_{AB} is set to 1 in Case 5 of Thurstone's law of comparative judgement).

The logistic distribution is approximately equivalent to the Normal distribution if $\sigma = 1.7/a$, as shown in Figures 3 and 4 below.

Figure 3 Probability density of the logistic and Normal (Gaussian) distributions

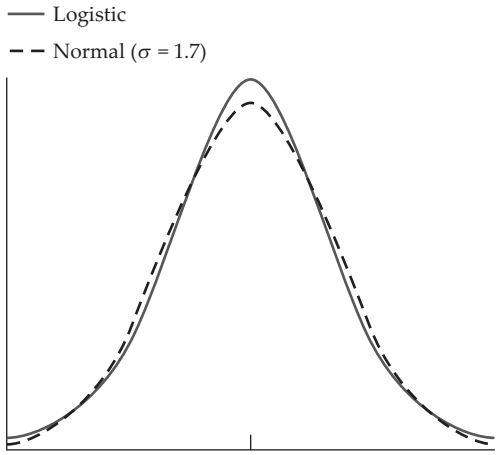
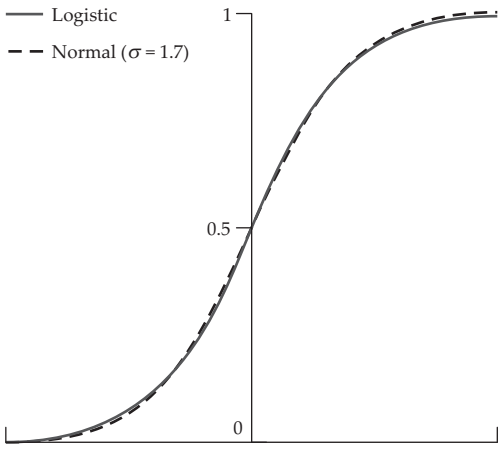


Figure 4 Cumulative probabilities of the logistic and Normal (Gaussian) distributions



Logistic models are widely used both in general categorical data analysis where equation (8) is known as the Bradley-Terry model (Bradley & Terry, 1952; Agresti, 1990); and specifically in Item Response Theory in the field of educational measurement, where equation (8) has the same form as that of Rasch's (1960) model for dichotomous items. The connections between the Rasch model and Thurstone's Case 5 have been explained in detail by Andrich (1978) and some of these issues will be picked up again later in this chapter.

Rearranging equation (8) into the same form as Thurstone's Case 5 equation (6) gives:

$$\ln\left(\frac{p(A>B)}{1-p(A>B)}\right) = \ln\left(\frac{p(A>B)}{p(B>A)}\right) = S_A - S_B \qquad (9)$$

Thus, rather than using the unit Normal deviate corresponding to the proportion of times A beats B to estimate the scale separation, this estimate is now based on the log of the ratio of wins to losses, that is, the log of the odds of success. The unit of the scale is known as a '*logit*', or 'log odds unit'. The logit scale is additive in the sense that the distance between any pair of objects A and B is the sum (difference) between each object and any other object C on the scale:

Log odds (A beats C) = $S_A - S_C$
Log odds (B beats C) = $S_B - S_C$

Subtracting:
Log odds (A beats B) = Log odds (A beats C) – Log odds (B beats C)

$= (S_A - S_C) - (S_B - S_C)$

$= S_A - S_B$

Thus Thurstone's model (in both its Case 5 and Rasch formulation) achieves the goal of sample-free calibration (Wright, 1977) in the sense that the estimate of the distance between A and B does not depend on which other objects they are compared with. The practical importance of this is that it is not necessary to have a completely crossed or balanced design of paired comparisons because the estimation process naturally handles missing data. An object's estimated measure will depend both on the proportion of times it has been the winner in its paired comparisons, but also on the quality (measures) of the objects it has been compared with. The precision (standard error) of the estimate will depend on both the number of comparisons involving the object, and on the information in each comparison – 'off-target' comparisons between objects widely spaced on the scale contribute less information than 'on-target' comparisons. The scale is equal-interval in the sense that a given logit difference between two objects has the same interpretation in terms of the probability of one beating the other in a paired comparison at all points on the scale.

The parameters are usually estimated by an iterative maximum likelihood procedure, which minimises the difference between the observed number of wins and losses and the expected number according to the model. It is worth noting that a measure cannot be estimated for any script that wins (or loses) every comparison it takes part in – we literally have no information about whether this script is just off the scale at the top (or bottom) end, or a long way off. Most software analysis programs will attempt to get round this problem by removing the script from the estimation process, then deriving a value by extrapolating from the measures that could be estimated.

2.3 Summary

In summary, the paired comparison method produces data that, when analysed according to the law of comparative judgement (Case 5), yield a value for each object on an equal-interval scale with an arbitrary origin and unit. The scale is equal-interval in the sense that the same distance between pairs of objects at different parts of the psychological continuum reflects the same probability of one 'beating' the other in a paired comparison. Equation (5) shows that the unit of measurement is the standard deviation of the distribution of discriminal differences, σ_{AB}, or 1.41σ where σ is the presumed constant discriminal dispersion of all the objects.

3 Paired comparisons in UK inter-board comparability studies

In an inter-board comparability study, the objects to be judged are scripts from the syllabuses of different awarding bodies that are intended to be 'comparable'. (Here a 'script' means the work of a candidate on all components contributing to the grading of a particular assessment, unless otherwise stated.) The judges are senior examiners or other experts nominated by the awarding bodies, henceforth called 'boards'. The paired comparison exercise usually takes place at a two-day meeting, where the judges repeatedly compare different pairs of scripts from the same grade boundary and decide which is the better, recording their judgement on a form. The different variables involved in the design and conduct of the exercise are described in some detail in section 3.2. The main advocate of the Thurstone paired comparison method

in comparability studies has been Alastair Pollitt, and most of the theoretical arguments and discussions about the application of the method have come from him and his colleagues at Cambridge Assessment[1]; and from Robert Adams and his colleagues at the Welsh Joint Education Committee (WJEC). The following section draws heavily on their work, mostly in the form of papers and presentations for technical seminars organised either by the QCA, or by the Joint Forum. See, for example, Pollitt & Murray (1996); Jones (1997; 1999; 2004); Bramley *et al*. (1998); Adams (1999); Pollitt (1999; 2004); Pollitt & Elliott (2003a; 2003b); Bramley (2005a).

The first part of a comparability study is called the 'specification (formerly syllabus) review'. One purpose of this is to provide a context for the expert judgements that take place in the second part of the study – the 'cross-moderation exercise'. In the specification review the judges consider the assessment-related material from each specification (content of syllabus, assessment structure and objectives, question papers and mark schemes). Various means of collecting, summarising and structuring the outcomes of this review have been tried over the years, for example getting the judges to write short reports, or fill in rating scales and questionnaires, or complete a Kelly repertory grid (see, for example, Elliott & Greatorex, 2002, or Chapter 5 of this book). For the purposes of this chapter, the importance of this part of the comparability study is that it gives the judges the opportunity to form an impression of the demand of the question papers – a factor that is highly relevant to the judgemental task to be carried out in the cross-moderation exercise.

The second part of a comparability study is the cross-moderation exercise where the judges from the different boards make judgements about the relative standards achieved by candidates in different assessments. This chapter concentrates on comparisons across boards within the same subject area in the same year, although the same methods have also been applied to comparisons over time.

The third part of a comparability study is the statistical analysis of grading outcomes. This does not involve expert judges and is the subject of Chapter 10.

3.1 Advantages of using Thurstone pairs methodology in cross-moderation exercises

Judges' internal standards cancel out

The Thurstone method has superseded the method known as 'ratification' or more informally 'home and away' (Elliott & Greatorex, 2002). The ratification method has been described in detail in Chapter 6.

In the ratification method teams of judges from different boards 'fix' their own board's standard in their mind, then make judgements about scripts from their own and other boards as to whether they are above, at or below that standard. The method thus relies on the judges having the same internal standard – or at least assumes the judges within a board will have the same internal standard (see Chapter 6). The main advantage of the Thurstone method is that the judges' internal standards 'cancel out' in the paired comparison judgement. This claim is made in

most discussions of the Thurstone method, but rarely spelt out, so for the sake of completeness it is explicated below.

Figure 5 Case 5 scenario for two scripts, A and B

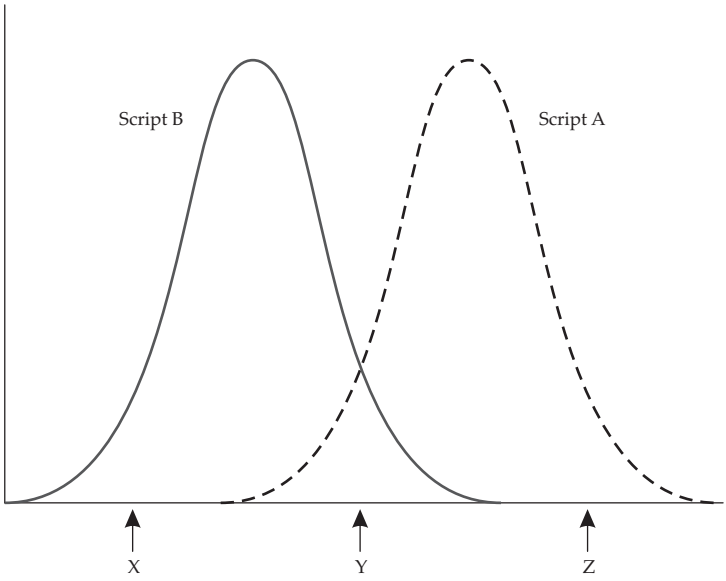


Figure 5 shows a Case 5 scenario, where the probability distributions apply to a group of judges. A and B are two scripts. If a particular judge has an internal standard located at point X on the psychological continuum then in the ratification method they would (probably) judge both A and B as being above the standard. If their internal standard was located at point Y they would (probably) judge script B as being below and script A as being above the standard, and if their internal standard was located at point Z they would (probably) judge both A and B as being below the standard. However, in the paired comparison method, the probability of judging A better than B depends only on the scale locations of A and B for all the judges.

This was shown mathematically by Andrich (1978), who included a parameter for what we might term judge 'severity'. A 'severe' judge is one for whom all scripts 'register' lower than average on the psychological continuum (i.e. they consistently think the quality of the script is lower than the other judges). Conversely a 'lenient' judge is one for whom all scripts register higher than average on the psychological continuum. In this formulation the scale location of the script is the mean of the discriminal processes across judges.

The 'discriminal process' D_{JA} for judge J on script A is:

$$D_{JA} = S_A - F_J + e_{JA} \qquad (10)$$

where S_A is the scale value of script A, F_J is the severity of judge J and e_{JA} is a random error term.

Correspondingly, the discriminal process D_{JB} for judge J on script B is:

$$D_{JB} = S_B - F_J + e_{JB} \quad (11)$$

Section 2 showed that according to Thurstone's model the outcome of the paired comparison judgement for an individual judge depends on the difference on the psychological continuum between D_{JA} and D_{JB}:

$$D_{JA} - D_{JB} = S_A - F_J - e_{JA} - (S_B - F_J - e_{JB}) = S_A - S_B - (e_{JA} - e_{JB}) \quad (12)$$

It can be seen from equation (12) that the judge severity F_J has 'cancelled out'.

It is interesting to note that the judge 'cancels out' *experimentally*, by virtue of the paired comparison design. This was the main point of Andrich's (1978) paper. It is not necessary to use the Rasch model to eliminate the judge *statistically* so in this sense the use of the Rasch model for Thurstone pairs analysis is 'gilding the lily' – more a matter of computational convenience than conceptual purity. However, there are other significant advantages of the Rasch framework in terms of estimation of parameters and analysis of residuals, which will be discussed later in this section. It is also worth noting that it would be possible to analyse data collected by the ratification method using the Rasch partial credit model (Wright & Masters, 1982), and thus obtain estimates of scale values for the scripts that are 'free' from the severities of the judges.

Forced choice removes problem of 'zero width'

The second weakness of the ratification method, as it tended to be used in cross-moderation exercises, was the potential for differences between the judges in how broadly they defined the 'standards equal' category. Since their judgements of 'lower standard', 'same standard' and 'higher standard' tended to be converted to –1, 0 and +1 for statistical analysis, this was referred to as the 'zero width' problem (e.g. Jones, 1997). On more than one occasion, authors have pointed out that a wide enough zero category virtually guarantees that no differences will be found amongst the boards, and some have implied that this might have been intentional (Pollitt & Elliott, 2003a)!

However, the 'zero width' drawback is not really a feature of the ratification method per se, but rather a consequence of giving the judges the option of using a middle category. It could be removed from the ratification method by forcing the judges to decide either 'above' or 'below' for all scripts; likewise the drawback could be introduced to the Thurstone method by allowing ties in paired comparison judgements. However, none of the comparability studies that have used paired comparisons has allowed the judges to make tied judgements. Various practical

strategies have been used to stop judges getting 'hung up' over a difficult judgement, such as telling them to toss a coin if they really cannot tell the scripts apart, or to make a note on their recording sheet of any comparisons they were unsure of.

The model handles missing data

As shown in section 2, if the data fit the Case 5 Thurstone model, the estimate of the separation between any two scripts does not depend on which other scripts they are compared with. This means that data can be missing in a non-random way without affecting the results. However, the precision (standard error) of each script's estimate depends largely on how many comparisons it has been involved in, so some effort needs to be made to ensure that each script is compared a similar number of times in total.

Fitting an explicit model gives insight into the data

Using the Rasch approach to analysing Thurstone pairs data means that the outcome of each individual comparison is explicitly modelled. The outcomes of the analysis can be considered in two parts: the 'model' and the 'misfit' (Pollitt & Elliott, 2003b). The 'model' part is the construction of the scale and estimation of the scale values of the scripts, along with the precision of these estimates (the standard errors). It is the features of the 'model' part that have been described in detail above. The 'misfit' part of the analysis investigates the degree to which the observed data fit the expectations of the model. Because the outcome of each individual comparison is explicitly modelled, it is possible to generate a 'residual' at the individual comparison level. This residual is the difference between the observed outcome (1 or 0, corresponding to win or loss) and the expected outcome (calculated from equation (8) using the estimated script parameters). The diagnostic potential of analysing these residuals has been highlighted by Pollitt (1999; 2004) and Pollitt & Elliott (2003a; 2003b). These residuals can be standardised and aggregated in various ways and used to investigate different questions of interest. For example, residuals can be aggregated for each judge to investigate judge misfit. Judges with a high value of misfit have tended to make more 'surprising' judgements than the other judges, that is, they have occasionally (or frequently) rated a low script above a high one on the scale. This can be interpreted as indicating that this judge has a different conception from the other judges about what makes a script better or worse.

Similarly, residuals can be aggregated for each script to indicate the extent to which the judges agreed on its location in the scale. Scripts with a high value of misfit have been involved in more surprising judgements than the other scripts and thus might be investigated to see if they contain anything unusual in terms of candidate performance.

If the surprising outcomes come mainly from paired comparisons involving scripts from the judge's own board then we have some evidence of *bias*. In practice, in most studies judges have not compared scripts from their own board in order to remove the possibility of bias. Other sub-sets of the matrix of residuals (for example male judges judging the work of male candidates) can be aggregated to investigate other forms of bias hypothesised by the analyst.

Finally, at a more mundane level, individual comparisons with a large residual might indicate an error by the judge in recording their judgement on the record sheet, or an error in data entry prior to analysis.

The data collection design is flexible

Since the analysis can cope with non-random missing data (see earlier), the collection of data can be very flexible – tailored to the needs of the particular situation. Because the comparative judgements are all independent, data can be accumulated as needed. The investigator is thus not (in principle) bound by a prior design. At any point more data could be collected and added to the existing data to recover from accidents, to replace an unacceptable judge, to add extra scripts, to resolve conflicts, to investigate possible bias or to deepen the study of any aspect that proves especially interesting. To date, this flexibility has not been fully capitalised on, due to the organisational constraints involved in getting around 20 judges in the same place at the same time. But with the advent of technology for making scanned images of scripts routinely available it is possible to imagine a distributed paired comparison exercise taking place online, with on-the-fly calibration guiding the selection of pairs of scripts to be presented to each judge (Pollitt, 2004).

3.2 Issues in designing paired comparison exercises

Many of the issues that arise and decisions that have to be made in planning and executing a paired comparison exercise are not unique to the Thurstone method, but will be discussed briefly below in terms of the Thurstone/Rasch conceptual framework. Appendix 1 summarises the 'design' features of inter-board comparability studies that have used Thurstone paired comparisons in the cross-moderation strand.

Judge selection

It is obviously important that the judges are appropriately qualified and capable of making the judgements. The validity of the whole exercise depends on the judges sharing, to a certain extent, the same conception of what makes a script better or worse. Whilst the paired comparison method allows for differences in absolute severity between the judges it does require that the underlying latent trait, or psychological continuum, is the same. It is also necessary for validity that the features of the scripts that influence the judges in making their decisions are related in the right way to the intentions of the assessment designers. One operational definition of question validity used in research on question design is:

> A question can only be valid if the students' minds are doing the things we want them to show us they can do.
>
> Ahmed & Pollitt (2001)

Clearly, by extension, the outcome of a judgemental exercise can only be valid if the judges are basing their judgements on the extent to which the students' minds have done the things the examiners wanted them to show us they could do!

There is therefore a limited pool of appropriately qualified expert judges. Those selected are usually senior examiners of the specifications in question from the different boards, who will usually have played a part in the awarding meeting that set the grade boundaries on the examination, and some of whom may have even set the questions. A typical comparability exercise involves around ten to twenty judges, usually two to three from each board. This is in notable contrast to Thurstone's own work where the number of judges was usually over 200! However, the main issue is whether the number of judgements per script is sufficient to allow the scale to be defined with reasonable precision. This is discussed in section 3.3.

Some studies have used 'independent' judges (with no affiliation to a particular board). One potential advantage of this is that it could add credibility to the results in the eyes of the general public (Bardell *et al.*, 1978, cited in Jones, 1999). The Thurstone pairs method is particularly suitable for the inclusion of independent judges because, as mentioned above, they would not need to have internalised a particular boundary standard in order to make their judgements. They would, however, need sufficient expertise to conceptualise the trait being assessed in the same way as the other judges, or their judgements would misfit. Forster & Gray (2000) found that independent judges were no more likely to produce misfitting judgements than board-affiliated judges.

However, Appendix 1 shows that in most of the more recent studies, there have in fact been no independent judges. The usual practice is for judges not to make paired comparison judgements about scripts that have come from their own board. In principle this restriction is probably not necessary, since it would always be possible to investigate the extent of 'home board bias' by analysing residuals (as in Pollitt & Elliott, 2003a). However, given that there is in practice not enough time for each judge to make all possible paired comparisons, it seems sensible to remove the potential for home board bias by design.

Script selection

Assessment or component?

The first choice that needs to be made is whether judgements are going to be made at the level of the whole assessment, or on individual components of the assessment. There are arguments in favour of each. On the one hand, the ultimate aim of the exercise is to confirm (or deny) that the boards' overall standards are in line, which suggests that judgements should focus on the assessment as a whole. Indeed the practice in most inter-board comparability studies has been to make judgements at the level of the overall assessment, as shown in Appendix 1. On the other hand, within each assessment the grade boundaries are set at the individual component level in the awarding meeting. The component grade boundaries are then aggregated to form the boundaries for the assessment as a whole. It is arguable, therefore, that the judges are more experienced at making valid judgements at the level of the component, and that this is where the standards 'really' reside. It has been suggested that future studies could make comparisons at unit level (Jones, 2004), but it seems that this is currently difficult in terms of time and cost.

There may be practical difficulties in bringing together the whole work of a candidate if coursework or oral components were involved in the assessment. It also becomes more difficult to find work exactly at a particular grade boundary as the effective raw mark scale increases (which it does by aggregating to assessment level).

On, or around the boundary?
The next decision to be taken is whether to use scripts exactly on the grade boundary, or covering a range of marks around the grade boundary. The former choice was necessary for the ratification method because that essentially involved asking judges whether work known to be at the boundary from a different board was perceived to be at the standard of boundary level work from their own board. However, it is by no means necessary for the paired comparison method, and it is argued in section 3.4 that there are good reasons for using a range of scripts around the boundary, and even in reconceptualising the comparability task as one of comparing whole mark scales rather than specific points on those scales (the grade boundaries).

Nevertheless, inter-board comparability studies have aimed to use scripts exclusively at the boundary marks. Appendix 1 suggests that the majority of studies have achieved this aim, but in practice it has not always been possible because of the difficulty of locating suitable scripts.

Composite scripts
Once the decision has been taken (as it usually is) to restrict scripts to those at the overall assessment grade boundary then if sufficient scripts cannot be located it is possible to create 'composite' scripts by putting together components from different candidates. These scripts are usually referred to in the reports as coming from 'pseudo-candidates'. Indeed composite scripts are sometimes used anyway in order to create a 'balanced' profile of performance (see below). The impact of composite scripts on the judgement process is not known, although there is some evidence that judges report finding them harder to assess (e.g. Arlett, 2002; Guthrie, 2003).

Balanced performance
It is deemed preferable to use scripts displaying a 'balanced' performance, which is usually taken to mean low variability of individual component totals, or of section or question totals within a component. In other words, an effort is made to avoid scripts that contain a mixture of questions or sections with very high marks along with other questions or sections with very low marks. Of course, there are many more ways to achieve an 'unbalanced' score profile than a balanced one, and the truly 'balanced' candidate is probably very unrepresentative of all the candidates! Scharaschkin & Baird (2000) reported that balanced and unbalanced scripts were judged differently in terms of their grade-worthiness, and that there were differences across subjects in this effect. This has also been reported in some of the inter-board comparability studies (e.g. Gray, 2000).

It is interesting to consider whether the concept of 'balance' could be further extended to include 'misfitting' score profiles (those containing a higher proportion of unexpectedly good answers to difficult questions or unexpectedly poor answers to

easy questions). It is quite possible for such a script to appear 'balanced' in the Scharaschkin & Baird sense of containing an even profile of raw marks across questions and sections, and yet it is easy to imagine it causing problems for the judges.

Cleaning of scripts

In some of the earliest studies using Thurstone pairs methodology (these were not inter-board comparability studies) the scripts involved were 'cleaned' of total marks and sometimes also of marker annotations and individual question mark totals (Bramley *et al*., 1998). This was to avoid the judges basing their judgements on simply adding up the marks. It could be argued that this is only a potential problem if within-assessment comparisons are to be made, and Appendix 1 shows that none of the inter-board studies have involved such comparisons. However, it does seem reasonable to assume that the presence or absence of marks and annotations might have an effect on how the judges make their judgements. None of the inter-board studies has used 'cleaned' scripts, so this is a potential area for future research.

Phrasing of the judgemental task

It is obviously of great importance to know how the judges actually make their judgements! It is equally obviously difficult to determine this. One way is by simply asking them what features of performance influenced them, as done by, for example, Adams & Pinot de Moira (2000), Fearnley (2000) and Edwards & Adams (2002). The problem with this of course, which is not unique to cross-moderation exercises, is that it is merely an instance of the general psychological problem of the validity of introspection or self-report in understanding the true causes of a judge's behaviour.

There is research suggesting that in some conditions self-reported explanations of behaviour are post-hoc justifications aimed at making the behaviour seem rational; and that human judgement in general is subject to many biases and unconscious influences (see, for example, Slovic & Lichtenstein, 1971; Nisbett & Wilson, 1977; Laming, 2004; Leighton, 2004). A preferable method might be to use Kelly's repertory grid technique (see Chapter 5) to discover what the judges perceived to be the salient constructs in the scripts they were comparing, and to relate these to their paired comparison judgements. This Thurstone and Kelly combined approach was used with some success by Pollitt & Murray (1993). A third method (as yet untried in this context) would be to carry out a controlled experiment, systematically varying the features of scripts involved in the comparisons.

In any event, the way the task is explained and phrased to the judges is presumably relevant. It was appreciated early on (Jones, 1997) that the judgements needed to be made quickly in order to complete a reasonable proportion of the possible judgements within the time available. In D'Arcy (1997), in the study comparing linear and modular mathematics A level the judges completed around 26–31% of the total possible comparisons. The judges in the study comparing linear and modular biology A level only managed to complete about 10% (A boundary) and 19% (E boundary) of the possible comparisons (see Appendix 2). Instructions in more recent

studies have thus emphasised the need for a quick, or impressionistic judgement. This is more in keeping with the spirit of Thurstone's method but it raises issues about the validity of the task, which will be discussed further in section 3.4. In terms of the judgement itself most studies have tended simply to ask the judges to decide which of the pair of scripts was 'better'. Some studies have used 'higher quality', or 'better in terms of attainment' or 'better in terms of performance'[2].

However, the need for the whole exercise in the first place arises from the fact that the different boards have different specifications and question papers. The judges are really therefore being asked to judge which performance is better, *taking into account any differences in the perceived demands of the questions (and specifications)*. This more complex question is sometimes left implicit, since the cross-moderation exercise follows the specification review, the aim of which is to encourage the panel to focus on similarities and differences in content and demand (amongst other things) in the assessment materials from each board. However, sometimes it has been spelt out more explicitly, as in the quotation below from the instructions to judges in the comparability study on GCE AS chemistry (Greatorex *et al*., 2002):

> The judgment is an impressionistic 'holistic' judgment as there is insufficient time for more deliberated judgments. In other words do not consider marks awarded, instead quickly read the whole script through, make a mental allowance for differences in specific questions and decide on balance, on the basis of all comparable material which script you feel 'has the edge'.
>
> Greatorex *et al.* (2002)

Allocation of scripts to judges

The general procedure in the inter-board studies has been to avoid within-board comparisons, and to avoid judges making judgements about scripts from their own board. The exceptions have been the comparability studies involving Vocational Certificate of Education (VCE) units (Arlett, 2002, 2003; Guthrie, 2003), where only three boards were involved.

There have been two general procedures for allocating pairs of scripts to judges – the more common one has been to provide the judges with a record sheet showing the entire set of possible comparisons (except for within-board comparisons), ask the judges to cross out the comparisons which they are 'ineligible' to make (i.e. those involving scripts from their own board), then for each new paired judgement to keep one script from the previous comparison and select a new one, never using the same script for more than two successive judgements. The judges are responsible for ensuring that they cover a representative proportion of the set of all possible judgements (i.e. they do not over-represent particular boards or scripts). This method (or variants of it) has broadly been followed in Gray (2000) and others (see Appendix 1). It is a pragmatic solution to the problem of scripts being unavailable if a specific schedule were given to every judge – because judges will inevitably work at different speeds. It is also probably the easiest to organise.

However, it does have the drawback of probably violating one of the assumptions of the Thurstone pairs method – that each paired comparison is independent of the others. If the same script is involved in consecutive pairs of judgements then it is highly likely that features of it will be remembered from one comparison to the next. Of course, this means that the judge will probably not need to read it twice (or at least, spend so much time reading it on the second comparison) so this does seem likely to allow more comparisons to be made in the time available.

The second approach has been to supply the judges with a unique individual schedule of paired comparisons at the level of the board (not individual script within board), ensuring that scripts from the same board are never used in successive comparisons. In some studies the judges have chosen an available script within the set for their assigned board, and in others the administrative staff have handed out specific scripts. This approach (or variants of it) has been followed in Adams & Pinot de Moira (2000) and others (see Appendix 1). This more rigorous approach is probably preferable.

Appendix 2 shows that the judges in studies using this method do seem to have made a lower proportion of the possible comparisons than the judges in studies that used the more pragmatic method, but it is not possible to attribute this entirely to the allocation method – first, because it is not always possible to determine from the reports how much time was available for the comparisons in each study, and second because different subjects produce different quantities of material for the judges to read.

Timings

Most studies have aimed to allow from two-and-a-half to five minutes per judgement. Some have tried to enforce this with a bleeper, set for a longer time interval at the start of the exercise then reducing the time when the judges have become familiar with the task. A comment from the judges, which has been recorded in more than one report, is that they find the bleeper irritating!

There is no particular rationale for this time limit – it has evolved through practice as allowing at least enough time for both scripts to be read through briefly (and thus varies depending on the subjects being compared). As Appendix 2 shows, even with this time restriction, the percentage of possible comparisons which are actually made is often relatively small.

3.3 Analysis of results

There has been some variability in how different authors have chosen to present the results of analysing their paired comparison data. Appendix 3 summarises some of the different features, discussed below.

Frequency table

This is the raw data upon which the analysis is based, and as such can give a very useful summary of both the experimental design (which scripts were compared, and how many times) and of the results (how many times each script won and lost). An

example is shown in Table 1. If the design is fully crossed (all possible comparisons made by all the judges) then a simple table of proportion of wins and losses is probably nearly as informative (and easier to communicate!) than the results of the Rasch analysis. Appendix 3 shows that some of the published reports have included such a table – ideally it would always be included.

Script measures

All the reports have included either a table or a plot of script measures, and some have included both. The most common type of plot shows all the scripts on a vertical

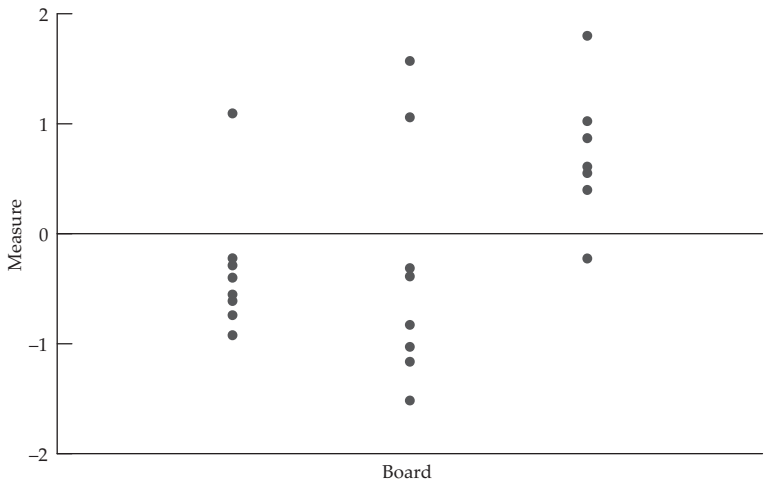
Table 1 Example frequency table from a paired comparison exercise (Edwards & Adams, 2002). Entries show the number of times a row script beat a column script.

		AQA	**CCEA**	**EDEXCEL**	**OCR**	**SQA**	**WJEC**
		12345	**12345**	**12345**	**12345**	**12345**	**12345**
AQA	1	00000	14102	10001	10000	10012	22001
	2	00000	01103	01220	00000	10041	00001
	3	00000	00011	02111	00201	01221	20101
	4	00000	02202	11411	01000	01001	10000
	5	00000	20001	00011	00000	00001	10000
CCEA	1	00012	00000	20110	00001	22000	11200
	2	00000	00000	00002	10000	00003	10000
	3	01010	00000	00011	00000	00000	00001
	4	01002	00000	02132	01100	01202	02100
	5	01001	00000	01031	00000	01121	00000
EDEXCEL	1	00011	11210	00000	00100	20121	00010
	2	00010	10010	00000	00010	10001	21001
	3	01110	01210	00000	00010	10011	00000
	4	10001	10102	00000	00000	01101	00000
	5	10210	11010	00000	00000	10001	00001
OCR	1	20001	21110	11312	00000	21010	21102
	2	10111	12122	11211	00000	42140	11003
	3	12102	11101	02111	00000	11200	22120
	4	21113	10101	43003	00000	10111	10021
	5	11112	22302	21010	00000	12130	11021
SQA	1	03001	11220	04112	00000	00000	00201
	2	10312	01021	00222	00111	00000	10010
	3	00002	02110	02020	00000	00000	00120
	4	00111	21001	10210	00000	00000	10001
	5	01120	01101	00000	00010	00100	00002
WJEC	1	00010	11002	01100	00001	20201	00000
	2	14010	31201	21003	10100	10020	00000
	3	12110	21201	40412	10000	12102	00000
	4	22011	12310	12131	01021	00111	00000
	5	21111	21100	01031	00000	01201	00000

scale with each script labelled. This gives a good visual impression of the differences between the scripts. A more detailed plot was chosen by Fearnley (2000) and Arlett (2002, 2003) which spread the scripts out along the horizontal axis and also included 95% upper and lower bounds based on the standard errors of the script estimates. The most recent report (Jones *et al*., 2004) used a bar chart instead of plotting points, and included a bar for the mean script measure for each board (but did not include standard errors).

None of the reports has yet used what is perhaps the most obvious chart (one used by Thurstone in his paper on the seriousness of different crimes, Thurstone, 1927d) – showing the spread of script measures on a separate vertical column for each board (but not necessarily identifying the individual scripts to avoid cluttering the chart). An example is shown in Figure 6, using data from Arlett's (2003) study at the A boundary in VCE health & social care.

Figure 6 Script measures by board for the 'A' boundary in Arlett (2003)



Such a chart allows relatively easy visual comparison of differences within and between boards.

Standard errors and separation reliability

Few of the reports have included the standard errors (SEs) of the script measures, which is a surprising omission, not because these values are of much interest in themselves, but because they allow an assessment of the extent to which the exercise has created a measurement scale. That is, it is useful to report on the extent to which differences between the scripts are due to 'true' differences in scale value, as opposed to measurement error. Several indices have been suggested for this (see, for example, Wright & Stone, 1979; Fisher, 1992), including the separation index, G, and the separation reliability (analogous to Cronbach's Alpha).

These indices are both calculated via the observed standard deviation (SD) of the script measures, and the root mean square error (RMSE) of the script SEs (effectively the 'average' measurement error). First, the 'true' SD of the script measures is calculated from:

$$(\text{True SD})^2 = (\text{Observed SD})^2 - (\text{RMSE})^2 \qquad (13)$$

Then the separation index G is calculated as the ratio of the 'true' spread of the measures to their average error, that is:

$$G = (\text{True SD}) / \text{RMSE} \qquad (14)$$

The separation reliability is defined like reliability coefficients in traditional test theory, as the ratio of true variance to observed variance, that is:

$$\text{Reliability (Alpha)} = (\text{True SD})^2 / (\text{Observed SD})^2 \qquad (15)$$

For example, from the data given in Fearnley (2000) we can calculate for the scale created at the A boundary: G = 4.50, Alpha = 0.95. This shows that only 5% of the variability between the script measures was due to measurement error. These are high values for the separation indices and help justify any later statistical tests. It is important to quote these figures as part of the context for interpreting both tests of fit (the lower the separation indices, the lower the power of tests of fit) and conventional significance tests for differences between board means (which allow for sampling error but not measurement error).

The issue of SEs was raised by Adams (1999) who was concerned by the small number of comparisons for each individual script pairing. He simulated a large number of data sets based on the script measures from the Pritchard *et al*. (2000) comparability study in GCSE English at the C (higher tier) boundary and obtained empirical estimates of the SEs which were comparable to those reported in Fearnley (2000) and Arlett (2002; 2003). He was also concerned about the implications of measurement error in the script estimates for the validity of the statistical tests of significant differences between board means and his suggested approaches are discussed below under the heading of 'interpreting differences between the boards'. He simulated some data based on random judgements (a 50% probability for each outcome in each paired comparison) but did not calculate the separation indices for this simulated data, which ought to have been close to zero.

Wood (1978) showed that random dichotomous data can fit the Rasch model quite well, which emphasises the need for separation indices to confirm the reliability of the scale, rather than relying on indices of model fit. However, this concern about reliability led to the best designed and most informative of all the inter-board paired comparison studies to date – the exercise on GCSE religious studies reported in Jones *et al*. (2004). This study was unique in both ensuring that all judges made all 'legitimate' comparisons (i.e. those not involving their own board, or any within-

board pairings), but more significantly in that there was a complete replication at each boundary with a different set of judges. The script measure estimates obtained from each replication were plotted against each other showing high correlations (> 0.8) giving an empirical demonstration of the reliability of the scale.

Fit of data to model

Most of the reports have indicated the proportion of misfitting judgements, which has usually been at or below the proportion (5%) that would be expected by chance, using a criterion value of 2 for the absolute value of the standardised residual. None of the reports has presented the usual Rasch fit statistics for the scripts or judges, but several of them have indicated that these statistics were examined. The mathematics study reported in D'Arcy (1997) reanalysed the data without a misfitting judge and noted that it had little effect on the outcome. The studies reported by Arlett (2003) and Guthrie (2003), where judges did make comparisons involving scripts from their own board, did contain 'home board' bias analyses of the kind advocated by Pollitt, but did not find anything significant and did not report any statistical details.

Interpreting differences between the boards

Most reports have presented the mean of the script measures for each board. Some (e.g. D'Arcy, 1997; Jones *et al*., 2004) also showed what the interpretation of these differences was in terms of the probability of the average script from board X winning a comparison against the average script from board Y. All the studies shown in Appendix 3 have reported the result of some kind of statistical test for differences between the board means. Most often this has been an ANOVA, occasionally repeated t-tests, and once (in Pritchard *et al*., 2000) a t-test of whether each mean board measure was significantly different from zero (the origin of the scale – set by default to be the mean of all the script measures).

Most authors have (rightly) been extremely cautious in interpreting the results of these statistical tests. For example, Adams & Pinot de Moira (2000) interpreted a significant result as:

> … unlikely to have arisen by chance; instead it may be concluded there are underlying differences among the syllabuses… An explanation for this may be underlying differences in grading standards among the syllabuses.
>
> Adams & Pinot de Moira (2000)

They were careful to point out that the scale created reflects the judges' preferences and that if they had been influenced in their judgements by extraneous factors such as handwriting then conclusions about differences in grading standard between the boards would not be valid. Many other authors have also been aware of this point and have tried to collect some feedback from the judges about how they made their decisions. See section 5 for further discussion of this issue.

A second reason to be cautious about the results of these tests is that they are essentially treating the scripts as random samples of boundary scripts from each board, and testing the hypothesis that there are no differences in the mean population boundary script judged measures, given the observed differences in the sample means. However, the design of the studies has ensured that the boundary scripts are *not* representative or random samples of all the boundary scripts – they are specifically chosen to have a balanced profile of performance, even to the extent of using composite scripts from pseudo-candidates. They are thus likely to be quite unrepresentative of the typical boundary script.

A final reason to be cautious is that the significance tests treat the script measures as constants – that is, they ignore the measurement error. This issue was first picked up by Adams (1999) who suggested aggregating the data prior to analysis by treating all the scripts from each board as a single script. This obviously would increase the number of comparisons per 'script' and hence the precision of the measure for each board. The resulting comparison between boards would simply be whether their measures were different from each other within the limits of measurement error. This would avoid the problem of testing hypotheses about sampling error mentioned earlier, but would obscure the differences between scripts from the same board (evident in Figure 6 and in all the other reports) and drastically reduce opportunities for investigating misfit.

A second possibility would be to combine the SEs for the individual scripts to obtain a measurement error for the board means. Assuming the errors are independent, the error in the total (sum) of each board's scripts is the square root of the sum of the squared SEs of its individual scripts. The measurement error in the mean, E, is therefore:

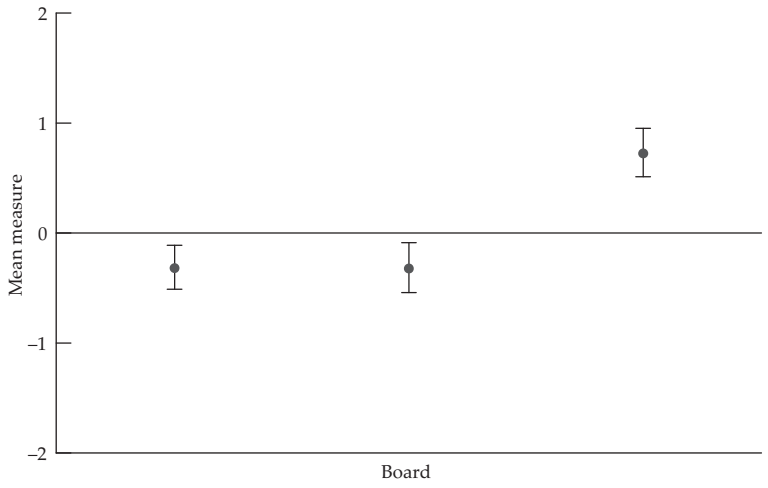
$$E = \frac{\sqrt{se_1^2 + \ldots + se_N^2}}{N} \qquad (16)$$

where N is the number of scripts from the particular board.

This could be presented graphically with 95% limits in a high–low chart as shown in Figure 7, where the data are taken from Arlett (2003).

A third possibility, and the most sophisticated statistically, is to analyse the data with a different model than the Rasch model (which can be considered as a special case of a general logistic regression model). John Bell has recommended this approach (in Greatorex *et al*., 2002), fitting various logistic regression models and arriving at a model that only included parameters for the board, plus those for 'aberrant' scripts with a value very different from the board mean. One advantage of this approach is that it takes into account the SEs associated with estimating the parameters when comparing differences between the boards. A further advantage is that it would be possible to develop models that involve fewer comparisons of a larger number of

Figure 7 Mean script measures by board (data from Figure 6) with 95% measurement error bars



scripts, with explicit modelling of script, judge and board parameters – but at the cost of moving away from the psychological and philosophical underpinnings of the Thurstone and Rasch method. It also increases the difficulty of communicating the results. This approach only appears to have been tried once so far, and there was only space for a brief explanation in the appendix of the report, but it may be an avenue for further investigation.

3.4 Evaluation of the use of paired comparisons in cross-moderation exercises

Psychological validity

The most obvious difference between the paired comparison method as used in cross-moderation exercises compared with those in Thurstone's original work is in the nature of the objects used for comparison. Even though Thurstone moved away from objects with both a physical and a psychological magnitude to objects with a purely psychological (or subjective) magnitude, the judgements could still be made very quickly: for example the judges might be asked to say which of two statements they agreed with more, or which of two handwriting samples they thought was better, or which of two statements reflected a more positive attitude towards religion. The scripts used in cross-moderation exercises, however, are the complete work of a candidate in the assessment and as such could take a long time to read in detail. Allowing five minutes for a paired comparison implies that judges are allowed about two-and-a-half minutes to read each script. This raises the question of whether the judges can apprehend enough of the relevant features of the scripts in such a short time to make a valid comparison between them.

In Thurstone's type of task, the objects could be perceived simultaneously, but in the cross-moderation task the 'discriminal process' of the first script has to be remembered when the second script is compared. Therefore there is an element of recall involved, and it is possible that features of the second script might 'interfere' with the memory of the first script. This raises the question of order effects in paired comparisons, which has not yet been investigated in this context.

Ironically, it is possible that the judges might resort to using an internal standard (this one is a good 'A') when making their judgements – even though one of the main benefits of the method is that in theory it removes the need to do this.

Model assumptions

As described in section 2, Thurstone derived five 'cases' of his law of comparative judgement, each making more assumptions than the last. The Rasch model used to analyse the data is analogous to his Case 5 law, which makes the most assumptions. Of these, the most questionable is the one that the discriminal dispersions of all the objects are equal. Thurstone clearly did not expect this to hold for any but the most simple stimuli. The scripts used in cross-moderation exercises are obviously far more complex than any used by Thurstone, so it seems naïve to expect this assumption to hold here.

Interestingly, inspection of equations (2) and (17) below shows that allowing discriminal dispersions in Thurstone's model to vary would be equivalent to allowing the scaling parameter *a* (see equation (8)) in the logistic form of the model to vary. This parameter is known as the 'discrimination' parameter in IRT modelling. It is inversely proportional to σ_{AB} in Thurstone's model – that is, the smaller the discriminal dispersions, the greater the discrimination, which makes intuitive sense.

$$X_{AB} = \frac{S_A - S_B}{\sigma_{AB}} \qquad (2)$$

$$\ln\left(\frac{p(A > B)}{p(B > A)}\right) = a_{AB}\left(S_A - S_B\right) \qquad (17)$$

The question of whether it is justifiable to use Thurstone's Case 5 law can now be seen to be analogous to the (much debated) question of whether it is justifiable to use a 1-parameter (Rasch) model rather than an IRT model containing more parameters. It is beyond the scope of this chapter to rehearse the arguments in this debate (see, for example, Goldstein, 1979; Divgi, 1986; Andrich, 1989; Wright, 1999). However, there seems to be consensus that the Rasch model is more robust and appropriate with small data sets, such as are produced in a cross-moderation exercise. Furthermore, once the reliability of the constructed scale has been verified with

separation indices, and misfitting data removed (if necessary), it is unlikely that using a more complex model would substantively alter the estimated measures of the scripts.

In order to make it possible to estimate the parameters satisfactorily with the relatively small amount of data available it would probably be necessary to reduce the number of parameters estimated in some other way, for example, by representing all the scripts from one board with a single parameter, as in the logistic regression model used in Greatorex *et al.* (2002). Such an approach shifts the philosophy of the exercise from that of Thurstone and Rasch (constructing psychological scales capable of yielding sample-free measurement) to the philosophy of statistical modelling (finding a model that optimises parsimony and variance explained).

However, it should be noted that the issue of discrimination does have an important bearing on the interpretation of results from the Rasch analyses, because each separate analysis creates its own logit scale with the discrimination parameter set to 1. This means that the 'unit' cannot be assumed to have the same meaning across analyses. Again, it is easiest to understand the implication of this by considering Thurstone's formulation (Figure 5). There is the same probability of script A beating script B in a paired comparison if they are a long way apart on the psychological scale, but with large discriminal dispersions, as there is if they are close together on the scale but with small discriminal dispersions. Different analyses define the unit in terms of the discriminal dispersion, but this unit will not have the same substantive meaning if the judges in one analysis are capable of more fine discriminations between the scripts than the judges in another analysis, or if one set of scripts is more conducive to fine discriminations than another (as might be hypothesised to occur between different subjects, for example).

Interpretation of results

Despite the caveats mentioned above, the cross-moderation exercises that have used Thurstone paired comparisons seem to have created valid measurement scales in the sense of being internally consistent (Fearnley, 2000; Arlett, 2002; 2003) and replicable (Jones *et al.*, 2004). This means that (in my opinion) the area where there is most scope for development is in the interpretation of the results, in the sense of what inferences can be drawn from differences in mean script estimates between the boards. In other words, there needs to be some way to determine an effect size, or to translate the differences in perceived quality of boundary scripts (in logits) into differences in grading standards (in marks).

If all the scripts from each board are exactly on the boundary, then the very fact that there is considerable variation in perceived quality within the scripts from each board would seem to provide a context for interpreting differences between the boards. For example, the mean difference between two boards could be expressed as a proportion of the SD of the measures within a board. But this would still only allow conclusions in terms of the psychological scale. What is really needed is a way to relate the psychological scale to the mark scale on which the original grade

boundaries were located. This point has been made before (Pollitt 1999; Pollitt & Elliott 2003a).

In order to achieve this, cross-moderation exercises should deliberately aim to use *ranges of scripts* around the boundary. This would have the following advantages:

- it would offer a means of validating the outcomes of the judgemental exercise (by comparing the script measures within each board with their original mark totals)
- it would allow the size of any differences that are found between boards to be quantified (albeit approximately) in terms of the original mark scale, and hence the importance of the difference to be evaluated
- it would reduce (ideally remove entirely) the need to use pseudo-candidates in order to create composite scripts exactly on the boundary mark.

This approach was used in studies comparing standards over time by Bell *et al*. (1998) and Bramley *et al*. (1998), although it was found that the range of marks needed to be quite wide in order to obtain a good relationship between mark and judged measure. Furthermore, on some of the occasions in inter-board studies where non-boundary scripts have been involved they have provided a useful validation of the judgement outcomes (Alton, personal communication; D'Arcy, 1997).

Nearly all the inter-board comparability studies using paired comparisons that have reported the views of the judges have mentioned that they found the task difficult, tedious and repetitive. One could take the view that they are being paid for their discomfort – but it might also be worth considering a method that removes some of the tedium, particularly if it can address some of the problems with the paired comparison method described above. The rank-ordering method described in the next section looks promising.

4 A rank-ordering method for investigating comparability

Given that the nature of the objects being compared (scripts) is such that paired comparisons are unlikely to be independent, and that the scripts take a long time to read, instead of asking judges to make repeated paired comparisons it might be advantageous to ask them to put sets of scripts into rank-order of perceived quality. It is then possible to extract paired comparison data from the rank-ordering in the form of '1 beats 2', '2 beats 3', '1 beats 3', and so on. This strategy was occasionally adopted by Thurstone himself, who wrote:

> The ideal form of the constant method is the method of paired comparison… but is also one of the most laborious experimental methods… Our present problem is to devise a plan whereby simple absolute rank order may be used as the experimental procedure with the advantages of the much more laborious constant method. Given the data for absolute rank order, we shall extract the proportion of judgments 'A is greater than B' for every possible pair of stimuli in the given series. These derived proportions will be used instead of the proportions that are obtained directly in the constant method.
>
> Thurstone (1931)

Rank-ordering has also been suggested by judges who have had to undergo a paired comparison exercise (Edwards & Adams, 2002) and by researchers who have had to organise one (Bramley *et al*., 1998)!

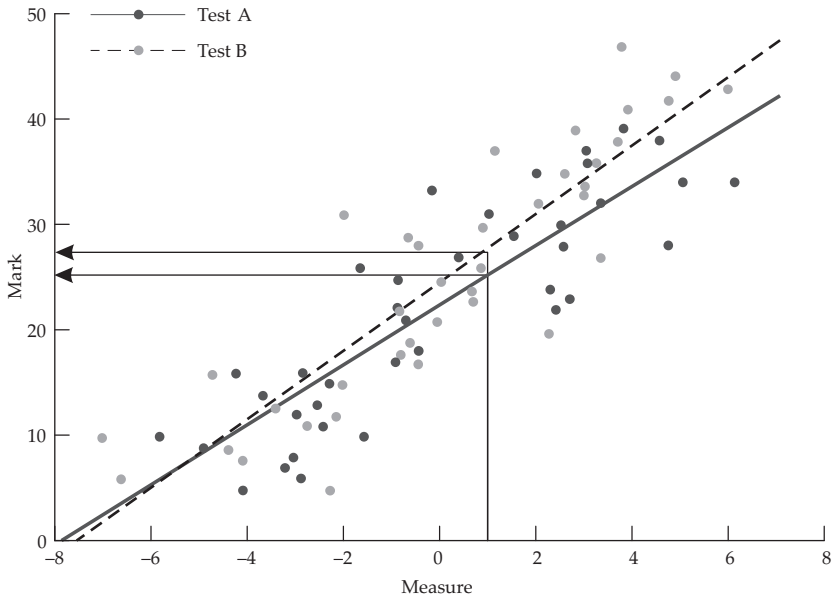
The application of rank-ordering methodology to the problem of standard maintaining within a subject over time has been described in detail in Bramley (2005a). The method has been used to map cut-scores from the Key Stage 3 English test from one year to another – effectively 'test equating by expert judgement'. The essentials of the method as it has been used to date are summarised very briefly below.

Scripts are selected from the two (or more) tests to be compared, such that the whole effective mark range is covered. Script total marks and question marks (if feasible) are removed from the scripts which are then photocopied as many times as necessary for the study. Packs of scripts are compiled containing scripts from both tests. In studies carried out to date, packs of ten scripts have been used containing five scripts from each test. The scripts within each pack can vary in both the range of marks covered, and in the degree to which the mark ranges from each test overlap or are offset. Each pack contains a unique selection of scripts, but there are many common scripts between the packs allowing the entire set of scripts to be 'linked'. Each judge is given a number of packs covering the whole mark range. Some effort is made to minimise the number of times each judge has to see the same script across packs, but some overlap is inevitable and improves the linking. Judges are asked to rank the scripts in each pack from best to worst, in terms of a holistic judgement of quality. Tied rankings are discouraged. Judges have access to the question papers and mark schemes in order to inform their judgements.

The ranked data are converted to paired comparison data prior to analysis with the Rasch model. The analysis places every single script in the study (i.e. covering the whole mark range of both tests) onto a single scale of perceived quality. Once the usual checks have been made for misfitting judgements, and scripts that have won or lost all their comparisons, the final output of the analysis is a graph that plots the script mark against the script measure for both tests separately, as shown in Figure 8.

The regression lines summarise the relationship between mark and measure and thus allow equivalent marks (in the sense of corresponding to the same perceived measure) on the two tests to be identified. For example, in Figure 8 a mark of 25 on Test A corresponds approximately to a mark of 28 on Test B. If the two tests had come from different boards, and the grade boundary marks were known, it would be possible to determine the boundary marks from board A that were equivalent to the boundary marks from board B, and hence to quantify (in marks) any difference in grading standard.

Figure 8 Example of test equating by expert judgement using the rank-ordering method



4.1 Evaluation of the potential of the rank-ordering method for cross-moderation exercises

The rank-ordering studies to date (UCLES, 2004; Bramley, 2005a; OCR & UCLES, 2005; Black & Bramley, in press) have just used single components of two assessments. The cross-moderation task involves the whole assessment from several boards. This difference should be borne in mind when considering the rank-ordering method as an alternative to paired comparisons.

Preparation

It is more time-consuming to prepare the materials for a rank-ordering exercise because of the need for cleaning scripts and photocopying them. (It is necessary to remove the marks because, unlike the typical inter-board cross-moderation study, comparisons are made between scripts from the same board and these comparisons should not be influenced by the rank-order according to total mark.) It is also a more complex task to design the allocation of scripts to packs. However, one benefit is increased control for the analyst over all aspects of the linking. Also, since all materials are prepared in advance, and judges work independently, it is possible for the exercise to be carried out postally, reducing the cost. Black & Bramley (in press) found that there was no difference in outcome between a rank-ordering exercise conducted postally and replicated in a face-to-face meeting.

Timing

A rank-ordering of 10 objects yields 45 paired comparisons for analysis[3]. In the studies to date, this has proved feasible in 30–45 minutes, giving a nominal time per comparison of less than a minute, compared to an average of around five minutes in the inter-board studies. There is therefore considerable scope for saving time with this method, even allowing for the fact that the scripts to be ranked in inter-board studies would require more reading time.

Judgemental task

Whilst the task is broadly the same as in a paired comparison exercise (a holistic judgement of overall quality), there is more scope for inter-judge differences in strategy for producing the ranking – for example some judges might read everything through very quickly to create a provisional ranking, then spend more time sorting out their preferred order for adjacent scripts. Others might work methodically through their pack, others might create separate rank orders for the scripts from the same assessment and then try to interleave them. Whether this affects the outcome has not yet been investigated.

Analysis

The requirement for each paired comparison to be independent is indisputably violated by creating the pairs out of a ranking. For example, if script A is ranked first, script B second and script C third then this creates the paired comparison outcomes 'A beats B', 'B beats C' and 'A beats C'. If these three scripts were to be compared in a 'true' paired comparison exercise it would be possible to obtain the inconsistent result 'A beats B', 'B beats C' and 'C beats A'. The ranking therefore constrains the possible set of paired comparison outcomes (Linacre, 2006). The more objects to be ranked, the greater the constraint. The ratio R of possible paired comparison outcomes from a ranking of N objects to the total set of possible outcomes is given in equation (18):

$$R = \frac{N!}{2^{\left(\frac{N(N-1)}{2}\right)}} \qquad (18)$$

However, in practice it is possible that the violations of local independence do not greatly affect the results. Bramley (2005a) showed that effectively the same set of script measures were produced by analysing the rankings with the Rasch partial credit model (PCM) as with the paired comparison model, but the latter appeared to create a more discriminating scale – that is, the separation (reliability) indices were artificially inflated. There is scope for further experimental investigation of the difference between measures created from rankings analysed as paired comparisons and measures created from a genuine paired comparison design.

For example, it may be that if the objects to be ranked are sufficiently far apart on the psychological scale then many of the possible outcomes in the denominator of equation (17) would be so unlikely as to have effectively a zero probability, making the constraint imposed by a ranking in practice much less than it seems in theory. Also, as mentioned in section 3.4, it seems quite plausible that even with a 'genuine' paired comparison design, when the objects being compared are as complex as scripts it is unlikely that each comparison will be truly independent because of memory effects, which might impose a 'virtual' constraint with the same effect as a ranking if the judges either consciously or unconsciously try to be self-consistent as they make their decisions.

Validity

The greatest benefit of using the rank-ordering method is the potential it creates to compare the rank-order of scripts by judged measure to the rank-order by mark – if this relationship is poor it casts doubt on the validity of the exercise. This can be visually assessed by considering the scatter about the regression lines in plots like Figure 8, or by considering indicators of fit such as R^2 or root mean square error (RMSE). Interestingly, the rank-ordering studies carried out to date have often shown that within a judge's pack the correlation between perceived quality and mark for scripts from the same test is often low, in fact sometimes negative – yet when the results are aggregated over the entire mark range for all judges the overall correlation between mark and measure is high (around 0.8 to 0.9).

Once a relationship between mark and measure has been established, it can be summarised by a linear (or non-linear) best-fit line, which effectively allows raw scores on one test to be mapped to the other test via the common construct of perceived quality. If several boards were involved then the actual grade boundary mark for one of them could be mapped to an equivalent boundary mark on the others. This equivalent mark could be compared with the actual boundary mark to find the difference in grading standard (in marks) between the boards, according to the expert judges.

In practice there might be several obstacles to overcome before this method could be used in an inter-board cross-moderation exercise. For example, it might not be feasible to compare the whole mark scale of a complete assessment for several boards in the time available. This could be overcome by using scripts a certain number of marks apart along the mark range, instead of at each mark. Alternatively, separate rank-ordering exercises could be carried out on ranges of marks around the key boundaries of interest.

In summary, the two key benefits of using a rank-ordering method would be to speed up the data collection, making the task less tedious and repetitive for the judges, and to create a way to relate the psychological scale of perceived quality to

the raw mark scale, allowing differences between the boards at the grade boundaries to be quantified in terms of raw marks.

5 Discussion

The purpose of inter-board comparability studies is to compare standards across the different awarding bodies. In making this comparison, it is important to distinguish between content standards and performance standards:

> Content standards refer to the curriculum [or syllabus/specification] and what examinees are expected to know and to be able to do… performance standards communicate how well examinees are expected to perform in relation to the content standards.
>
> Hambleton (2001)

The syllabus or specification review strand of a comparability study is an exercise comparing the content standards of the different specifications. From the point of view of teachers and pupils, it is probably true to say that the content standards are most effectively communicated by the question papers and mark schemes for the assessments. The specification review strand thus considers all this material (specifications, question papers and mark schemes) from each board.

It is important to note that judgements about content standards can often depend on values. For example, a mathematics specification including calculus might be seen as having 'higher' content standards than one which did not include calculus but did include matrices, if it were judged that calculus were more intrinsically demanding, or of more fundamental importance to a mathematician. Such a judgement would inevitably have a value-laden subjective element.

Performance standards relate to where the grade boundaries are set on a particular assessment – that is, how many marks are 'good enough' for a script to be worthy of a particular grade. Because question papers inevitably vary in difficulty, despite the best efforts of the paper setters, it is often necessary to set different grade boundaries on different papers to allow for differences in difficulty. The purpose of an awarding meeting is to set the boundaries at points on the raw mark scale that represent an equivalent performance standard to the boundaries set on previous assessments in the same specification. The content standards do not vary from year to year (unless the specification changes). The awarding meeting uses a variety of evidence in order to decide whether to make any adjustments for differences in difficulty. One of these pieces of evidence is expert judgement about the quality of work displayed in scripts on a range of marks around the putative boundary. Once the boundary has been set, scripts on the grade boundary (in conjunction with the question papers and mark schemes) therefore exemplify the performance standard.

In a cross-moderation exercise, boundary scripts from different boards are compared. The exercise is thus comparing the performance standards of the different boards. The extra complication in attempting to accomplish this task in a cross-moderation exercise (compared to in an award meeting) is that not only are the question papers

and mark schemes different, the specification is too. Thus the judges are expected to make judgements about relative *performance* standards in a context of possible differences in *content* standards.

I have argued elsewhere (Bramley, 2005b; 2006) that a psychometric approach provides a rational framework for tackling the problem of maintaining performance standards, even if it cannot provide an entirely satisfactory solution. On this approach, the performance standard can be conceived as a location on a latent trait (the psychological continuum representing the construct that the assessment is measuring). Pupils and test questions can also be located on the trait by analysing test performance data (item-level data of pupil marks on individual questions) with a measurement model such as the Rasch model. A question's location on the trait is referred to as its 'difficulty', and a pupil's location as their 'ability'[4]. If there are some common questions in two otherwise different tests that are testing the same construct then it is possible to link scores on one test to scores on the other in terms of both representing the same amount of ability.

The latent trait itself thus embodies the content standards – in the sense that the set of questions spread out across the latent trait would define the construct. Other questions measuring the same construct could also be calibrated on the same scale (provided there was a link via common items or pupils). The level of ability at the performance standard can be interpreted in terms of probability of success on questions at different points on the trait, providing a coherent framework for understanding both content and performance standards.

I would like to suggest that it is *only possible to compare performance standards if the content standards across the boards are similar enough for the different assessments to be considered to be measuring the same construct*. Then we could imagine calibrating the test questions from all the boards onto a single scale. Having done this, we might observe that the average location of the questions from different boards was at different points on the scale – in other words that their tests varied in difficulty. This would mean that a candidate with a given level of ability would get a lower raw score on the assessment from the 'difficult' board than from an 'easier' board. However, this would *not imply anything* about the performance standards of the boards! This important point has been made on several occasions by Robert Adams (e.g. Adams & Pinot de Moira, 2000; Jones, 1999):

> There is no obvious connection between the demand of the papers and grading standards because awarding intervenes. The most demanding papers can yield the most easily attainable grades if grade boundaries are set low enough; similarly higher grades can be difficult to obtain on easier tests if the boundaries are set high enough.
>
> Adams & Pinot de Moira (2000)

The way to use our imaginary calibrated scale to compare the performance standards of the different boards would be to use the measurement model to plot the relationship between raw score and ability on each of the boards' assessments. Then a difference in performance standard would be indicated if the raw score grade

boundaries set by the different boards corresponded to different abilities on the overall calibrated scale.

Whilst it is possible to carry out this calibration in the imagination, it is not possible to do it in reality because there are no common questions between the boards' assessments, and no common candidates. Thus the cross-moderation exercise can be conceived as an attempt to achieve this calibration via a link of 'common judges'. What do the judges need to be able to do in order to achieve this? It seems unlikely that their brains are running the kind of iterative maximum likelihood algorithms that the computer would employ if we had a common question or common candidate link! Given a script containing a set of responses to a set of questions with an associated mark scheme, they need to be able to perceive directly the location of the script (i.e. candidate ability) on the latent trait – the *same* latent trait as imagined above. In Thurstone's terms, the script must evoke a 'discriminal process' at the point on the latent trait corresponding to candidate ability. Whilst in Thurstone's type of experiment this was an immediate judgement, in the cross-moderation exercise it seems more plausible that it must be an aggregation of micro-judgements as the judge reads through the pairs of scripts. We expect them to compensate for lower performance when the questions are harder. That is, we want them to be able to compare better performance on easier questions with weaker performance on harder questions. We can imagine this being done by aggregating a series of micro-features from the individual questions and the candidates' responses to them according to the judge's own idiosyncratic weighting system. Studies which have collected feedback from the judges on how they made their decisions have usually found a variety of different features that different judges say that they pick up on – naturally these tend to be very subject-specific.

The purpose of the mark scheme, it might be argued, is to impose a standardised weighting onto specified features of the response in order to allow an explicitly observable numerical aggregation of question marks into a total score. This raw score of course depends on the difficulties of the questions. (Otherwise it would be possible to compare the performance standards of the boards by comparing the raw boundary marks expressed as a percentage of total mark available.) On a Rasch-calibrated latent trait there is a one-to-one (sigmoidal) relationship between raw score and ability for each particular test. In other words, scripts with the same raw score will imply the same ability measure. That is why it is important to check that the measure scale created in the paired comparison exercise corresponds to the (within-board) raw score scale[5]. It is easier to do this with the rank-ordering method than with the paired comparison method, as shown in Figure 8. When all the scripts in a paired comparison exercise are on the boundary mark then all that can be done is to compare the extent to which within-board differences in script measure compare with between-board differences in script measure. There should be no within-board differences for scripts on the same mark.

This verification of a within-board relationship between judged measure and raw mark is necessary to compare performance standards in the current system where grade boundaries are set on the raw mark scale. It is possible to envisage a situation whereby the judges' judgements are held to be *more* valid (i.e. to give a truer estimate of locations on the latent trait) than the locations derived from adding up marks

according to the mark scheme. In this utopia cross-moderation exercises (and, indeed, marking!) would become obsolete (Pollitt, 2004).

Adams has expressed doubts that the judges are capable of allowing for differences in demand of question papers when comparing performances:

> It is, after all, the central conundrum of testing educational attainment: the relative merits of easier tasks done well and harder tasks done moderately.
>
> Adams & Pinot de Moira (2000)

Nonetheless, all cross-moderation exercises (regardless of the method they use) do depend on the ability of judges to do this. Experimental work could try to verify whether or not they are capable of it, in situations where it is possible to compare the judged equating of two mark scales with a statistical equating based on common items or persons. Good & Cresswell (1988) found considerable disparity between judgemental and statistical equating in tiered examination papers, but some recent work using the rank-ordering method (Black & Bramley, in preparation) has shown that judges did agree reasonably well with the outcomes of an awarding meeting in aligning the mark scales on two tiers in a GCSE English exam. If they can make this kind of adjustment for difficulty (between two tiers intended to be at a different level of difficulty), then there is some hope that they might be able to allow for the (presumably lesser) differences in difficulty between assessments at the same level from different boards.

6 Conclusion

The paired comparison method of constructing psychological scales based on human judgements is well established in psychometric theory and has many attractive features that have led to its adoption as the preferred method in inter-board comparability studies. Its main theoretical advantages are the experimental elimination of the internal standards of the judges when estimating scale locations, and the fitting of an explicit statistical model that allows investigation of residuals for script and judge misfit, and for various sources of bias. Its main practical advantage is the simplicity and flexibility of the design, which can allow for non-random missing data. Most of the more problematic issues discussed in this chapter arise from its application to the particular context of investigating examination comparability. The fact that scripts are complex objects which take a relatively long time to read, means that the comparison judgement is not based on an immediate impression, in contrast to Thurstone's work. Order effects and memory effects are likely to come into play, making the assumption of independence between judgements, required by the statistical analysis model, seem rather implausible. The paired comparison method is also time consuming and tedious for the judges, a drawback that can be remedied to some extent by using the rank-ordering method to collect the data.

However, the most serious problem in the cross-moderation exercises carried out to date has not been with the method but with the design of the studies, which, by only including scripts on the grade boundaries, have not allowed differences between

boards in terms of mean scale location to be related to the raw mark scales of the different examinations. This has made it impossible to assess the importance (for example in terms of implied changes to grade boundaries) of any differences discovered. It has only been possible to draw very tentative conclusions about perceived differences in quality of average borderline scripts, based on significance tests of dubious appropriateness. Both the paired comparison method and especially the rank-ordering method could easily address this shortcoming in future studies, by deliberately involving scripts from a range of marks around each grade boundary.

Further areas for future research are in understanding better the psychological processes involved in the paired comparison judgements, and in discovering the features of the scripts that are most influential in determining the outcome. The validity of the cross-moderation exercise (regardless of which particular method is used) depends on achieving a match between the judges' collective perception of the trait of 'quality of performance', and the trait as intended by the question paper setters and instantiated in the mark scheme. This could be assessed by considering the within-board relationship between mark and trait location, but, again, this would need the design of the studies to be modified to include scripts on a range of marks, rather than on a particular boundary mark.

Finally, the prospect of the availability of high-quality scanned images of scripts will allow researchers to improve the design of studies and reduce logistical problems. It might be possible to carry out the exercise online, involving a larger number of judges and scripts. The data could then be analysed 'on-the-fly', allowing the allocation of pairs of scripts to be targeted so as to achieve the maximum information from each comparison, and for misfitting scripts (and judges) to be removed.

Endnotes

1. Cambridge Assessment is the brand name of the University of Cambridge Local Examinations Syndicate (UCLES), a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

2. These quotes have come from the actual instruction sheet given to the judge panel, if this was provided in the report. Otherwise, it has been taken from explanatory text within the report. It is possible that the explanatory text does not reflect the literal instruction given to the judges.

3. The number of different paired comparisons in a rank-ordering of N objects is $N(N-1)/2$.

4. Ability here does not imply innateness, or IQ, or potential. It is used in a neutral psychometric sense of location on the trait.

5. Perhaps ideally this should be the ability measure rather than raw score. However, the relationship between raw score and ability is linear over most of the raw score range.

References

Adams, R.M. (1999, November). *The Rasch model and paired comparisons data: Some observations*. Paper presented at a seminar held by the Research Committee of the Joint Council for General Qualifications, Manchester. Reported in B.E. Jones (Ed.), (2000), *A review of the methodologies of recent comparability studies*. Report on a seminar for boards' staff hosted by the Assessment and Qualifications Alliance, Manchester.

Adams, R.M., & Pinot de Moira, A. (2000). *A comparability study in GCSE French including parts of the Scottish Standard grade examination. A study based on the summer 1999 examination. Review of question paper demand, cross-moderation study and statistical analysis of results.* Organised by Welsh Joint Education Committee and Assessment and Qualifications Alliance on behalf of the Joint Forum for the GCSE and GCE.

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Ahmed, A., & Pollitt, A. (2001, September). *Improving the validity of contextualised questions*. Paper presented at the British Educational Research Association Annual Conference, Leeds.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 449–460.

Andrich, D. (1989). *Distinctions between assumptions and requirements in measurement in the social sciences*. In J.A. Keats, R. Taft, R.A. Heath & S.H. Lovibond (Eds.), *Mathematical and theoretical systems* (pp. 7–16). North Holland: Elsevier Science.

Arlett, S. (2002). *A comparability study in VCE health and social care, units 1, 2 and 5. A study based on the summer 2001 examination*. Organised by the Assessment and Qualifications Alliance on behalf of the Joint Council for General Qualifications.

Arlett, S. (2003). *A comparability study in VCE health and social care, units 3, 4 and 6. A study based on the summer 2002 examination*. Organised by the Assessment and Qualifications Alliance on behalf of the 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.

Bell, J.F., Bramley, T., & Raikes, N. (1998). Investigating A level mathematics standards over time. *British Journal of Curriculum and Assessment, 8*(2), 7–11.

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*.

Bradley, R.A., & Terry, M. (1952). The rank analysis of incomplete block designs: I. The

method of paired comparisons. *Biometrika*, *39*, 324–345.

Bramley, T. (2005a). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, *6*(2), 202–223.

Bramley, T. (2005b). Accessibility, easiness and standards. *Educational Research*, *47*, 251–261.

Bramley, T. (2006, March). *Equating methods used in key stage 3 science and English*. Paper presented at the National Assessment Agency technical seminar, Oxford.

Bramley, T., Bell, J.F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*, *25*(2), 1–23.

D'Arcy, J. (Ed.). (1997). *Comparability studies between modular and non-modular syllabuses in GCE Advanced level biology, English literature and mathematics in the 1996 summer examinations*. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

Divgi, D.R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, *23*, 283–298.

Edwards, E., & Adams, R. (2002). *A comparability study in GCE AS geography including parts of the Scottish Higher grade examination. A study based on the summer 2001 examination*. Organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications.

Edwards, E., & Adams, R. (2003). *A comparability study in GCE Advanced level geography including the Scottish Advanced Higher grade examination. A study based on the summer 2002 examination*. Organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications.

Elliott, G., & Greatorex, J. (2002). A fair comparison? The evolution of methods of comparability in national assessment. *Educational Studies, 28*, 253–264.

Fearnley, A. (2000). *A comparability study in GCSE mathematics. A study based on the summer 1998 examination*. Organised by the Assessment and Qualifications Alliance (Northern Examinations and Assessment Board) on behalf of the Joint Forum for the GCSE and GCE.

Fisher, W. (1992). Reliability statistics. *Rasch Measurement Transactions*, *6*(3), 238.

Forster, M., & Gray, E. (2000, September). *Impact of independent judges in comparability studies conducted by awarding bodies*. Paper presented at the British Educational Research Association annual conference, University of Cardiff.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, *5*, 211–220.

Good, F.J., & Cresswell, M.J. (1988). *Grading the GCSE*. London: Secondary Examinations Council.

Gray, E. (2000). *A comparability study in GCSE science 1998. A study based on the summer 1998 examination*. Organised by Oxford Cambridge and RSA Examinations (Midland Examining Group) on behalf of the Joint Forum for the GCSE and GCE.

Greatorex, J., Elliott, G., & Bell, J.F. (2002). *A comparability study in GCE AS chemistry: A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2001 examination*. Organised by The Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations on behalf of the Joint Council for General Qualifications.

Greatorex, J., Hamnett, L., & Bell, J.F. (2003). *A comparability study in GCE A level chemistry including the Scottish Advanced Higher grade. A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2002 examinations.* Organised by The Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations on behalf of the Joint Council for General Qualifications.

Guthrie, K. (2003). *A comparability study in GCE business studies, units 4, 5 and 6 VCE business, units 4, 5 and 6. A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2002 examination*. Organised by Edexcel on behalf of the Joint Council for General Qualifications.

Hambleton, R.K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jones, B.E. (Ed.). (1997). *A review and evaluation of the methods used in the 1996 GCSE and GCE comparability studies*. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

Jones, B.E. (Ed.). (2000). *A review of the methodologies of recent comparability studies*. Report on a seminar for boards' staff hosted by Assessment and Qualifications Alliance, Manchester.

Jones, B.E. (2004). *Report of the JCGQ research seminar on issues related to comparability of standards, 3 December 2003*. Internal Research Paper RC/264. Manchester: Assessment and Qualifications Alliance.

Jones, B., Meadows, M., & Al-Bayatti, M. (2004). *Report of the inter-awarding body comparability study of GCSE religious studies (full course) summer 2003*. Assessment and Qualifications Alliance.

Laming, D. (2004). *Human judgment: The eye of the beholder*. London: Thomson.

Leighton, J.P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, *23*(4), 6–15.

Linacre, J.M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement*,

7(1), 129–139.

Nisbett, R., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Oxford Cambridge and RSA Examinations and University of Cambridge Local Examinations Syndicate. (2005). *NAA KS3 English 2005 draft level setting report*. Report to National Assessment Agency.

Pollitt, A. (1999, November). *Thurstone and Rasch – Assumptions in scale construction*. Paper presented at a seminar held by the Research Committee of the Joint Council for General Qualifications, Manchester. Reported in B.E. Jones (Ed.), (2000), *A review of the methodologies of recent comparability studies*. Report on a seminar for boards' staff hosted by the Assessment and Qualifications Alliance, Manchester.

Pollitt, A. (2004, June). *Let's stop marking exams*. Paper presented at the annual conference of the International Association for Educational Assessment, Philadelphia.

Pollitt, A., & Elliott, G. (2003a). *Monitoring and investigating comparability: A proper role for human judgement*. Cambridge: Research and Evaluation Division, University of Cambridge Local Examinations Syndicate.

Pollitt, A., & Elliott, G. (2003b). *Finding a proper role for human judgement in the examination system*. Cambridge: Research and Evaluation Division, University of Cambridge Local Examinations Syndicate.

Pollitt, A., & Murray, N.L. (1993). What raters really pay attention to. Language Testing Research Colloquium, Cambridge. Reprinted in M. Milanovic & N. Saville (Eds.), (1996), *Studies in language testing 3: Performance testing, cognition and assessment*. Cambridge: Cambridge University Press.

Pritchard, J., Jani, A., & Monani, S. (2000). *A comparability study in GCSE English. Syllabus review and cross-moderation exercise. A study based on the summer 1998 examinations*. Organised by Edexcel on behalf of the Joint Council for General Qualifications.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

Scharaschkin, A., & Baird, J. (2000). The effects of consistency of performance on A level examiners' judgements of standards. *British Educational Research Journal*, *26*, 343–357.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, *6*, 649–744.

Thurstone, L.L. (1927a). Psychophysical analysis. *American Journal of Psychology*, *38*, 368–389. Chapter 2 in L.L. Thurstone (1959), *The measurement of values*. Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1927b). A law of comparative judgment. *Psychological Review*, *34*, 273–286. Chapter 3 in L.L. Thurstone (1959), *The measurement of values*. Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1927c). A mental unit of measurement. *Psychological Review*, *34*, 415–423. Chapter 4 in L.L. Thurstone (1959), *The measurement of values*. Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1927d). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, *21*, 384–400. Chapter 7 in L.L. Thurstone (1959), *The measurement of values*. Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1927e). Equally often noticed differences. *Journal of Educational Psychology*, *18*, 289–293. Chapter 5 in L.L. Thurstone (1959), *The measurement of values*. Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1927f). Three psychophysical laws. *Psychological Review*, *34*, 424–432. Chapter 6 in L.L. Thurstone (1959), *The measurement of values*. Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1931). Rank order as a psychophysical method. *Journal of Experimental Psychology*, *14*, 187–201. Chapter 10 in L.L. Thurstone (1959), *The measurement of values*. Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1945). The prediction of choice. *Psychometrika*, *10*, 237–253. Chapter 13 in L.L. Thurstone (1959), *The measurement of values*. Chicago, Illinois: University of Chicago Press.

Thurstone, L.L. (1959). *The measurement of values*. Chicago, Illinois: University of Chicago Press.

University of Cambridge Local Examinations Syndicate. (2004). *KS3 English 2004 draft level setting report*. Report to Qualifications and Curriculum Authority.

Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology*, *31*, 27–32.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97–116.

Wright, B.D. (1999). Fundamental measurement for psychology. In S.E. Embretson & S.L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65-104). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B.D., & Stone, M. (1979). *Best test design*. Chicago: MESA Press.

Appendix 1 Design of inter-board comparability studies (part 1 of 2)

Author(s)	D'Arcy	D'Arcy	Adams & Pinot de Moira	Gray	Pritchard, Jani, & Monani	Fearnley	Edwards & Adams
Year published	1997	1997	2000	2000	2000	2000	2002
Year of exam	Summer 1996	Summer 1998	Summer 1999	Summer 1998	Summer 1998	Summer 1998	Summer 2001
Subject	Biology	Mathematics	French	Science	English	Mathematics	Geography
Level	A level (linear v modular)	A level (linear v modular)	GCSE	GCSE	GCSE	GCSE	AS (+ Scottish Higher)
Number of boards	1	4	7	6	6	6	6
Number of components	8 (3L + 5M)	14 (4L + 10M)	3 (Listening, Reading, Writing)	4	3	? - included coursework	3 (2 for SQA)
Number of boundaries	2 (A and E)	2 (A and E)	3 (A & C Higher, C Found.)	3 (A & C Higher, C Found.)	3 (A & C Higher, C Found.)	2 (C Inter. A Higher)	2 (A and E)
Number of judges	12	12	20	18	18	17	17
Any independent judges?	No	No	Yes (6)	Yes (6)	Yes (6)	Yes (5)	No
Number of scripts per board per boundary	A: 10M, 6L E: 6M, 6L	6	5	5	5	5	5
Scripts cleaned of marks?	No?	No?	No?	No?	No?	No?	No?
All scripts exactly on the boundary?	Yes L, No M	No	Yes	Yes	Yes	No?	Yes?
Assessment or component level judgements?	Assessment	Assessment	Component	Assessment	Assessment	Assessment	Assessment
Any composite scripts?	Yes?	Yes (in modular)	No	No?	No	No?	Yes
How were pairs assigned?	?	Planned rotation of scripts	Random scripts from prescribed pairs of boards: never see the same board in consecutive judgements	Random selection by judges, keeping one and swapping the other each time	Random selection by judges, keeping one and swapping the other each time	Random selection by judges, keeping one and swapping the other each time	Random scripts from prescribed pairs of boards: never see the same board in consecutive judgements
Ties allowed?	No	No	No	No	No	No	No
Within board comparisons?	None within scheme	No	No	No	No	No	No
Judge own board's scripts?	Yes?	Yes?	No	No	No	No	No

Abbreviations: L = linear syllabus, M = modular syllabus,
Found = Foundation tier, Inter = Intermediate tier

Appendix 1 Design of inter-board comparability studies (part 2 of 2)

Author(s)	Greatorex, Elliott & Bell	Arlett	Edwards & Adams	Greatorex, Hamnett, & Bell	Arlett	Guthrie	Jones, Meadows & Al-Bayatti
Year published	2002	2002	2003	2003	2003	2003	2004
Year of exam	Summer 2001	Summer 2001	Summer 2002	Summer 2002	Summer 2002	Summer 2002	Summer 2003
Subject	Chemistry	Health and social care	Geography	Chemistry	Health and social care	Business studies (GCE), Business (VCE)	Religious studies
Level	AS (+ Scottish Higher)	VCE	A level (+ Scottish Advanced Higher)	A Level (+ Scottish Advanced Higher)	VCE	A2, VCE	GCSE
Number of boards	6	3	6	6	3	5 (A2), 3 (VCE)	4
Number of components	3 (2 for SQA)	3 (Units 1, 2 & 5)	3	Range from 2 to 6	3 (Units 3, 4 & 6)	3	2
Number of boundaries	2 (A and E)	1 (E)	2 (A and E)	2 (A and E)	2 (A and E)	2 (A and E)	2 (A & C)
Number of judges	16	8	18	17	9	20	11, replication 12
Any independent judges?	No	No	No	No	No	No	No
Number of scripts per board per boundary	5	10	5	5	8	5	5
Scripts cleaned of marks?	No?	No?	No?	No?	No?	No?	No?
All scripts exactly on the boundary?	Yes?	Yes?	No	No?	Yes?	Yes?	No
Assessment or component level judgements?	Assessment	Assessment	Assessment	Assessment	Assessment	Assessment	Assessment
Any composite scripts?	?	Yes	Yes	Yes	Yes	Yes	No
How were pairs assigned?	Random selection by judges, keeping one and swapping the other each time	Random selection by judges, keeping one and swapping the other each time	Random scripts from prescribed pairs of boards: never see the same board in consecutive judgements	Random selection by judges, keeping one and swapping the other each time	Random selection by judges, keeping one and swapping the other each time	Random selection by judges, keeping one and swapping the other each time	Random prescribed order: never see the same script in consecutive judgements
Ties allowed?	No	No	No	No	No	No	No
Within board comparisons?	No	No	No	No	No	Only for GCE v VCE	No
Judge own board's scripts?	No	Yes	No	No	Yes	Yes	No

Appendix 2 Percentage of possible judgements made in inter-board comparability studies

Author(s) of report	Published	Year of exam	Subject	Level	Boundary	Number of comparisons	Number of possible comparisons	%
D'Arcy	1997	Summer 1996	Biology AEB	A level	Ai2	69	720	10%
					Ai1	68	720	9%
					E	81	432	19%
D'Arcy	1997	Summer 1998	Mathematics	A level	A	811	2,592	31%
					E	684	2,592	26%
Adams & Pinot de Moira	2000	Summer 1999	French, Listening	GCSE	C Found.	812	8,400	10%
					C Higher	809	8,400	10%
					A Higher	846	8,400	10%
Adams & Pinot de Moira	2000	Summer 2000	French, Reading	GCSE	C Found.	826	8,400	10%
					C Higher	755	8,400	9%
					A Higher	860	8,400	10%
Adams & Pinot de Moira	2000	Summer 2001	French, Writing	GCSE	C Found.	804	6,000	13%
					C Higher	718	6,000	12%
					A Higher	727	6,000	12%
Gray	2000	Summer 1998	Science	GCSE	C Found.	1,606	5,250	31%
					C Higher	1,675	5,250	32%
					A Higher	1,743	5,250	33%
Pritchard, Jani & Monani	2000	Summer 2001	French, Writing	GCSE	C Found.	804	6,000	13%
					C Higher	964	5,250	18%
					A Higher	907	5,250	17%
Fearnley	2000	Summer 1998	Mathematics	GCSE	C Inter.	2,173	4,875	45%
					A Higher	2,157	4,875	44%
Edwards & Adams	2002	Summer 2001	Geography	AS	A	537	4,250	13%
					E	718	4,250	17%
Greatorex, Elliott & Bell	2002	Summer 2001	Chemistry	AS	A	876	4,000	22%
					E	907	4,000	23%
Arlett	2002	Summer 2001	Health & Social Care	VCE	E	960	1,920	50%
Edwards & Adams	2003	Summer 2002	Geography	A level	A	525	4,500	12%
					E	652	4,500	14%
Greatorex, Hamnett & Bell	2003	Summer 2002	Geography	A level	A	525	4,500	12%
					E	1003	4,250	24%
Arlett	2003	Summer 2002	Health and Social Care	VCE	A	739	1,584	47%-
					E	929	1,584	59%
Guthrie	2003	Summer 2002	Business studies (GCE), Business (VCE)	A2, 1 VCE	G v G:A	536	5,000	11%
					V v V:A	350	1,500	23%
					G v V:A	647	7,500	9%
					G v G:E	565	5,000	11%
					V v V:E	287	1,500	19%
					G v V:E	557	7,500	7%
Jones, Meadows & Al-Bayatti	2004	Summer 2003	Religious Studies 1	GCSE	A	825	825	100%
					C	825	825	100%
Jones, Meadows & Al-Bayatti	2004	Summer 2003	Religious Studies 2	GCSE	A	900	900	100%
					C	900	900	100%

Abbreviations: i1 = Indicator 1, i2 = Indicator 2, Found. = Foundation tier, Inter. = Intermediate tier, G = GCE, V =VCE

Indicator 1 is the term used to refer to the assessment grade boundary derived from weighted aggregation of the component grade boundaries. Indicator 2 refers to the assessment grade boundary derived from weighted averaging of the cumulative percentages of candidates at the component grade boundaries.

Appendix 3 Presentation of paired comparison results in inter-board comparability studies (part 1 of 2)

Author(s)	D'Arcy	D'Arcy	Adams & Pinot de Moira	Gray	Pritchard, Jani, & Monani	Fearnley	Edwards & Adams
Year published	1997	1997	2000	2000	2000	2000	2002
Year of exam	Summer 1996	Summer 1998	Summer 1999	Summer 1998	Summer 1998	Summer 1998	Summer 2001
Subject	Biology	Mathematics	French	Science	English	Mathematics	Geography
Level	A level (linear v modular)	A level (linear v modular)	GCSE	GCSE	GCSE	GCSE	AS
Frequency table	No	No	Yes	No	Yes	No	Yes
Plot of script measures	Yes	Yes	No	Yes	Yes	Yes	No
Table of script measures	No	No	Yes	No	Yes	Yes	Yes
Standard errors	No	No	No	No	No	Yes	No
Scale separation /reliability	No	No	No	No	No	No	No
Script misfit	No	No	No	No	No	Yes (not presented)	No
Judge misfit	No	Yes	No	Yes	Yes	Yes (not presented)	No
% misfitting judgements	No	E 0.7%, A 2.5%	No	FC 2.1%, HC 4.4%, A 2.9%	A 3.1%, HC 4.1%, FC 3.5%	C 3.6%, A 4.5%	No
Size of difference between means	Yes	No?	Yes	No	Yes	Yes	Yes
Sig. of difference between means	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Statistical test	t-test (2 groups)	ANOVA	ANOVA	t-tests	ANOVA + t-test of difference between mean and zero	ANOVA + t-test of difference between mean and zero	ANOVA + t-test of difference between group means

Abbreviations: Sig. = Significance, FC = Foundation tier C boundary, HC = Higher tier C boundary, G = GCE, V = VCE

Appendix 3 Presentation of paired comparison results in inter-board comparability studies (part 2 of 2)

Author(s)	**Greatorex, Elliott & Bell**	**Arlett**	**Edwards & Adams**	**Greatorex, Hamnett & Bell**	**Arlett**	**Guthrie**	**Jones, Meadows & Al-Bayatti**
Year published	2002	2002	2003	2003	2003	2003	2004
Year of exam	Summer 2001	Summer 2001	Summer 2002	Summer 2002	Summer 2002	Summer 2002	Summer 2003
Subject	Chemistry	Health and Social Care	Geography	Chemistry	Health and Social Care	Business studies (GCE), Business (VCE)	Religious studies
Level	AS	VCE	A level	A level	VCE	A2, VCE	GCSE
Frequency table	No	No	Yes	No	No	No	No
Plot of script measures	Yes	Yes	No	Yes	Yes	Yes	Yes (bar chart)
Table of script measures	Yes	Yes	Yes	Yes	Yes	Yes	No
Standard errors	No*	Yes	No	No	Yes	No	No
Scale separation and reliability	No	No	No	No	No		
Script misfit	No*	Yes (not presented)	No	No	Yes (not presented)	No	No
Judge misfit	No	Yes (not presented)	No	No	Yes (not presented)	No	No
% misfitting judgements	A 5.4%, E 3.2%	E 2.7%	No	A and E 0.009%	A 4.1%, E 3.8%	G v G:A 5.2% G v G:E 4.4% V v V:A 4.6% V v V:E 4.2% G v V:A 4.8% G v V:E 4.8%	No
Size of difference between means	No	No	Yes	Yes	No	Yes	Yes
Sig. of difference between means	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Statistical test	t-tests + logistic regression	ANOVA	ANOVA + t-test of difference between group means	t-tests	ANOVA + post hoc comparison of differences between group means	ANOVA + post hoc comparison of differences between group means	t-tests

* Standard errors and indication of fit were given in the logistic regression output in Greatorex *et al.* (2002).

COMMENTARY ON JUDGEMENTAL METHODS

Sandra Johnson

An alternative to ratification and paired comparisons

The ratification and paired comparison methodologies, described in Chapters 6 and 7 respectively, are at extremes in terms of technical sophistication and complexity. Yet studies that have employed these techniques have shared a common and critical weakness. This is that even where evidence has emerged of relative severity or leniency in boards' grading standards, it has not been possible to indicate by how much and in what direction boundary marks should be moved to bring standards into line (see Edwards & Adams, 1997; Adams & Pinot de Moira, 2000; Pritchard *et al*., 2000; Greatorex *et al.*, 2002; Guthrie, 2003).

While ratification studies have been able to identify which board(s), if any, could be considered to have the 'right' standard at the boundary under investigation, it has not been possible to quantify the sizes of other boards' deviations from this standard, nor to indicate appropriate remedial action in terms of how boundary marks might need to be adjusted. In paired comparison studies it has not even been possible to identify those boards whose standards could be considered appropriate, given that scrutineers have by design made *relative* judgements about script quality rather than *absolute* judgements about merited grades. Extending paired comparison studies to embrace scripts from a range of marks around boundaries, rather than confining them to scripts at specific boundary points, might improve their ability to quantify the size of difference between different boards' standards. But this will not solve the problem of identifying which board, if any, represents the 'correct' standard (zero points on Rasch scales are not indicators of this), nor how boundary marks for other boards should be moved to rectify deviations from it.

An alternative methodology that potentially solves this problem is a 'distribution study'. Here, boards supply equal numbers of randomly selected scripts at each grade (not exclusively at grade boundaries), and scrutineers independently evaluate and award grades to each script within randomly sorted batches, thus creating new distributions in which new boundary marks emerge. Such studies were piloted more than 25 years ago (Johnson & Cohen, 1983; see also Cohen & Johnson, 1982; Johnson & Cohen, 1984; Johnson, 1989) in a formative evaluation of the cross-moderation methodology as then practised (see Bardell *et al*., 1978 and Forrest & Shoesmith, 1985, for reviews), and produced promising results. Funded by the Schools Council[1], the research aimed not only to identify weaknesses in the way that cross-moderation techniques had been used in the past – these were well-known already – but also to offer recommendations for design improvement that would render cross-moderation

more 'fit for purpose' in the future. A brief reference to this work is included in Chapter 6, but with no discussion of the findings or implications.

Several issues were addressed in the evaluation project, in the form of a number of research questions. Was the assumption underpinning traditional cross-moderation studies tenable, *viz*. that individual examiners could 'carry' and hence 'represent' their boards' grading standards in such contexts? Could scrutineers consistently apply standards – their own or their boards' – when making grading judgements? Could cross-moderation provide clear evidence of differences in boards' grading standards? Where apparent differences in standards emerged, would they necessarily be uniform across the grade range? How much confidence might be attached to scrutineers' perceptions of standards differences – in other words, how technically reliable could cross-moderation outcomes be? How might future cross-moderation studies be better designed to produce more useful outcomes than had been possible in the past?

Three studies were carried out within the formative evaluation, and their results interpreted within the framework of generalizability theory (Cronbach *et al*., 1972; Cardinet & Tourneur, 1985; Shavelson & Webb, 1991; Brennan, 1992; 2001). Generalizability theory is an extension of classical test theory, and uses variance components derived from ANOVA modelling to estimate measurement reliability. A generalizability study first requires the identification of observable factors that can be assumed or suspected to affect the dependent variable, the dependent variable in this case being scrutineers' grade judgements. An appropriate ANOVA design is then identified with which to investigate factor effects; here a mixed-model factorial design was appropriate, with examiners, nested within examining boards, independently judging scripts nested within board batches, with every examiner judging every script (a 'script' being the total available evidence of a candidate's work – objective test results, responses to structured examination papers, oral test results, etc.). Relative influences on the dependent variable are then quantified in the form of variance components: boards' standards, scripts, interactions between examiners and scripts, etc. The component information is in turn used to calculate 'generalizability coefficients' – ratios of linear combinations of adjusted components that indicate the technical reliability of various pieces of evidence, such as between-board grade differences. Finally, a 'what if?' analysis offers predictions of reliability when features of the current design are changed – such as increasing the number of scripts scrutinised or the number of scrutineers involved.

The studies focused on one or other of three different subjects and levels, *viz*. CSE physics, O level French and A level mathematics, and each involved three senior examiners from each of three participating boards. For each subject, relatively large samples of scripts were randomly selected from across the grade range in each board, using disproportionate stratified sampling to ensure equal numbers of scripts from within each grade: 84 or 98 scripts in total from each board in each subject, comprising 14 scripts from within each of the six or seven grade bands at the relevant examination level. The scripts were randomly distributed into two sets: one set to be evaluated by 'immediate impression' during a 2-day residential meeting, the other

for evaluation in a later at-home exercise, in which scrutineers would mark the scripts before awarding grades, using the relevant boards' mark schemes (to see whether a deeper understanding of the qualities being valued in different schemes might produce different grading judgements).

As is current practice, for their information and orientation the scrutineers were sent syllabuses and examination papers for review before they arrived at the residential meeting. During the meeting, scrutineers independently evaluated every script from every board, awarding what they considered to be an appropriate grade. Scrutineers worked through the batches in a given order (their own board last), but no constraint was put on the order in which they judged individual scripts within a batch. To ensure that all scripts were indeed scrutinised by every scrutineer, the scripts, which were original material, were assigned and labelled with unique but arbitrary identifiers, and each individual was given a grade recording sheet, which listed all the scripts within their batches. The same procedure applied to the follow-on at-home exercise, but this time the examiners were provided with the boards' mark schemes and were asked to mark each script with the appropriate mark scheme before making a grade judgement.

Evaluation study findings and implications

One firm finding was that there was no evidence that scrutineers from any one board 'shared' that board's grading standards. In other words, there was as much variation in grading standards among the three scrutineers 'representing' a particular board as there was among the scrutineers in general. No individual examiners, however experienced, could be assumed to have been able to reproduce in these studies the grade distributions originating from their own boards' routine grading exercises. The previous assumption that examiners could singly 'carry' their boards' standards in cross-moderation exercises was shown to be untenable.

Interestingly also, there was no evidence of any 'home board' effect. This is where scrutineers tend to rate scripts from their own board more highly than those from other boards. Presumably the requirement to judge individual scripts within batches rather than batches as entities, as in ratification studies, served to eliminate this possibility. That said, there was evidence in all three studies of a different, less interpretable, scrutineer-by-board interaction. On the evidence of script judgements, different scrutineers tended to judge the different batches, or boards, more or less leniently than others on average. This could be explained by differences in the qualities of subject performance tapped in the different boards' examinations that different scrutineers valued to different extents – differences in personal judging criteria, which the scrutineers were unfortunately not able to articulate. Or it could have reflected prior prejudices on the part of some or all scrutineers about the relative quality of the different boards themselves. If the latter, this would have serious implications for the interpretive value of paired comparison studies, since decisions about relative script quality might to an extent reflect general perceptions about differences between boards; this threat to validity will be exacerbated when ties are not allowed.

There was also evidence, in the form of statistically significant 'interaction' effects between scrutineers and scripts, of a lack of *general* consistency in the grading judgements of individual scrutineers. Such interaction effects will contribute to the phenomenon of 'misfitting judgements' in paired comparison studies, and could in principle threaten the validity of the Rasch model for such comparability exercises.

In two of the subjects, physics and mathematics, clear evidence emerged of overall differences in the boards' grading standards – differences that reached statistical significance at the 1% and 5% levels respectively, with generalizability coefficients of around 0.8 in each subject. Interestingly, while no reliability statistics have been offered in ratification study reports (for example, Edwards & Adams, 1997), generalizability coefficients *could* have been produced for these studies also, since they too have been based on repeated measures designs (a feature that also, moreover, demands F-ratios and not the typically quoted chi-square statistics for significance testing).

In French, no perceived board difference was evident. But the generalizability coefficient was very low, and there was significantly more inter-scrutineer variation here than in the other two subjects, posing a serious challenge for data interpretation. Was there genuinely no difference in board standards in O level French? Or were the scrutineers too variable in their judgements for any difference to emerge? The fact that not all of a candidate's work could be presented for evaluation must be relevant here, oral tests and objective papers being represented by marks only. This issue of missing evidence continues to be a problem today in many subjects.

To the critical question of whether or not cross-moderation studies might be able to provide guidance about how specifically to bring divergent standards together, the answer has to be in the affirmative, provided only that we are prepared to accept that the most valid indicator of 'true' standards lies in the joint judgements of experienced board examiners. By ranking the physics and mathematics scripts in order of original aggregated examination marks, and calculating 'majority votes' on the basis of the nine independent grade judgements provided by the scrutineers, it was possible to see, quite literally, how boundary marks should appropriately have been moved this way or that, and by how much, to achieve standards equivalence. *Every* board would have to have taken action at more than one boundary to achieve parity in standards, and the actions, typically one- or two-mark shifts, would have varied both in direction and magnitude from one boundary to another (see Johnson & Cohen, 1983, Appendices 4 and 6). The weakness in the ratification and paired comparison methods is their inability to provide this kind of information.

Conclusion

Challenges will continue to be faced by those engaging in grade comparability investigations, given the context of diversity that gives rise to the endeavour in the first place. And there will always be doubts about the extent to which the combined judgements of individuals, whether board examiners or others, reflect 'the truth' in

terms of a 'national standard'. Despite the difficulties, grade comparability must continue to be investigated.

It is not clear why the formative evaluation findings had so little impact in the UK examining world. In particular it is difficult to understand why ratification studies, with all their demonstrated weaknesses, continued to feature. Could the untimely demise of the Schools Council be a factor? Certainly, the planned continuation of the research was abandoned by default when the Council was closed. Was the relatively unfamiliar methodology not sufficiently well explained in the report for others to adopt? Or was the methodology well understood, but its model assumptions rejected? Was the problem the absence of any user-friendly software with which to carry out generalisability analyses? Or was it simply that the number of examining boards in operation in the mid-1980s, along with the workload that each examiner would be required to accept if all relevant boards were to be involved in any single subject study, precluded any possibility of more widespread application?

The number of examining boards is markedly lower today than it was 25 years ago, the methodology – a special case of multilevel modelling – is more familiar, and G-study software is now readily available.[2] The formative evaluation confirmed the potential of cross-moderation as a fit-for-purpose grade comparability tool. All that remains is for the necessary follow-on research to be carried out to refine the methodology, and for more informative study designs to be adopted in the future.

Endnotes

1 The evaluation report was published just before a government announcement that the Schools Council was soon to be replaced by the School Curriculum and Assessment Authority, another of QCA's predecessors.

2 This has been rectified in the interim with the availability of GENOVA (see Brennan, 2001, Appendix F), and more recently still with the more versatile and more user-friendly EduG (downloadable as freeware from www.irdp.ch/edumetrie/logiciels.htm).

References

Adams, R.M., & Pinot de Moira, A. (2000). *A comparability study in GCSE French including parts of the Scottish Standard grade examination. A study based on the summer 1999 examination. Review of question paper demand, cross-moderation study and statistical analysis of results.* Organised by Welsh Joint Education Committee and Assessment and Qualifications Alliance on behalf of the Joint Forum for the GCSE and GCE.

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.

Brennan, R.L. (1992). *Elements of generalizability theory* (2nd ed.). Iowa City: ACT Publications. (First Edition: 1983).

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer.

Cardinet, J., & Tourneur, Y. (1985). *Assurer la mesure*. Berne: Peter Lang.

Cohen, L., & Johnson, S. (1982). The generalizability of cross-moderation. *British Educational Research Journal*, *8*, 147–158.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York:Wiley.

Edwards, E., & Adams, R.M. (1997). *A comparability study in Advanced level English literature*. Cardiff: Welsh Joint Education Committee on behalf of the Joint Forum for the GCSE and GCE.

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.

Greatorex, J., Elliott, G., & Bell, J.F. (2002). *A comparability study in GCE AS chemistry: A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2001 examination.* Organised by The Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations on behalf of the Joint Council for General Qualifications.

Guthrie, K. (2003). *A comparability study in GCE business studies, units 4, 5 and 6 VCE business, units 4, 5 and 6. A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2002 examination*. Organised by Edexcel on behalf of the Joint Council for General Qualifications.

Johnson, S. (1989). Évaluation de la comparabilité des notations entry jurys d'examens. *Mesure et Évaluation en Éducation*, *12*, 5–22.

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

Johnson, S., & Cohen, L. (1984). Cross-moderation: A useful comparative technique? *British Educational Research Journal*, *10*, 89–97.

Pritchard, J., Jani, A., & Monani, S. (2000). *A comparability study in GCSE English. Syllabus review and cross-moderation exercise. A study based on the summer 1998 examinations*. Organised by Edexcel on behalf of the Joint Council for General Qualifications.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park: Sage.