CAMBRIDGE ASSESSMENT

Locating objects on a latent trait using Rasch analysis of experts' judgments

Tom Bramley

Paper presented at the conference "Probabilistic models for measurement in education, psychology, social science and health", Copenhagen, Denmark, June 2010.

Research Division
Cambridge Assessment
1, Regent Street
Cambridge
CB2 1GG

Bramley.T@cambridgeassessment.org.uk

www.cambridgeassessment.org.uk

UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

Introduction

Rasch models have been widely applied in the field of educational testing. If it is accepted that it is possible for several different tests to 'measure the same thing' then Rasch models provide an intuitively appealing framework in which to make inferences about the location of examinees and test items on the dimension / construct / attribute / variable / trait that tests designed to 'measure the same thing' are measuring. The advantages of Rasch models have been extolled by several authors, perhaps most prominently Ben Wright (e.g. Wright 1977; 1997; 1999).

Criticisms of the Rasch model (or more accurately, the uses to which it has been put, as few seem to dispute the elegance of the model) have been made on various grounds. In the 1970s and 1980s criticism focused on whether other models were 'better' in terms of fitting observed data more closely, and whether basing assessment on the Rasch model could compromise validity, leading to tests that sampled the domain of content and skills too narrowly, with a limited variety of item type, items being selected or discarded on a procrustean basis of model fit. (e.g. Goldstein, 1979; Divgi, 1986). Some of these criticisms were addressed as the variety of models in the Rasch family grew and the availability of software for fitting them increased.

More recently in the 1990s and 2000s a second wave of criticism has swept ashore as the more fundamental philosophical issues relating to psychological measurement have come under close scrutiny – for example the meaning of measurement in psychology (e.g. Michell, 1997; Maraun, 1998), the relationship between Rasch models and representational measurement theory (e.g. Kyngdon, 2008; Borsboom & Zand Scholten, 2008; Michell, 2008), and the relationship between different approaches to measurement and different philosophies of science (Borsboom, 2005). However, many of these criticisms have been directed more at the discipline of psychometrics as a whole than specifically at the use of Rasch models.

Whilst acknowledging the importance of these fundamental issues, this paper is written from the more pragmatic perspective of an assessment agency in England in the position of trying to ensure that its procedures are defensible and its outcomes are acceptable, both to the people who take the tests and to the wider community of stakeholders – parents, teachers, employers, regulatory bodies and the media. One type of outcome of particular interest is the distribution of grades awarded to examinees in a high-stakes examination of educational attainment in an academic subject (e.g. a General Certificate of Education (GCE) A level in Physics); or the proportion of passes and fails in a vocational examination (e.g. a Certificate of Professional Competence (CPC) in National Road Haulage[1]).

In both these situations, two statements are widely held to be true – in the sense that they operate as constraints on what assessment agencies can do:

1. Examinees with the same total score should receive the same grade (or pass-fail decision), i.e. within a test[2], the ranking by total score is equal to the ranking in terms of what the test is measuring.
2. Different versions of the same test can vary in difficulty, i.e. the ranking by total score from two or more tests does *not* necessarily give a fair ranking in terms of what the tests are measuring.

The first of these constraints means that it would be perceived as unfair to use a scoring model that resulted in examinees with the same total score receiving different grades (as might happen for example if item scores were weighted by discrimination). The second constraint means that

[1] See for example http://www.ocr.org.uk/qualifications/type/vrq/cpc/nrh_l3/index.html

[2] Almost all GCE and GCSE examinations involve aggregating scores on several units or components, which may carry different 'weight'. Grading decisions are usually made at the level of the unit/component, however, so the term 'test' in this paper should be taken as applying to these units/components.

it would be perceived as unfair to (automatically) set the grade boundaries[3] in the same place on the raw mark scales of different versions of the same test.

Most examinations are only available a limited number of times per year. For GCE A levels there are two sessions per year (in January and June). The high-stakes nature of the tests means that there is no item re-use, and cost and security concerns generally prevent any pre-testing. Each session's examination is therefore completely new. Setting the grade boundaries on the tests comprising successive versions of such examinations is therefore a practical problem that needs an acceptable and defensible solution. Both of the above-mentioned constraints motivate the use of the Rasch model as a tool for solving this problem, because:

– there is a sense in which it is accepted that items from different tests 'measure the same thing', so a latent trait model captures that intuition;
– in the Rasch model, the raw score is a sufficient statistic for estimating 'ability' (Andersen, 1977). (Constraint 1);
– in the Rasch model, the difficulty of a test is solely a function of the (varying) difficulty of the items comprising it (Constraint 2).

Analysing a matrix of person-item response data with the Rasch model results in a set of estimates of item difficulties and person abilities. It is natural to represent these as points on a line, with the line itself representing the trait that the test is measuring, as in Figure 1 below.

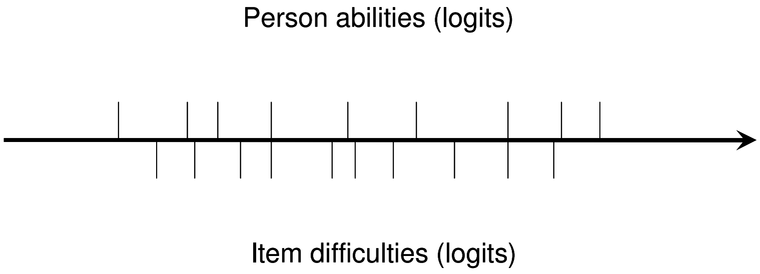

Figure 1: Representation of person ability and item difficulty parameters from the Rasch model for dichotomous items.

If the data fit the Rasch model, distances on the line between a person and an item correspond to the log of the odds of observing a particular outcome on any person-item encounter. Distances on the line between any two persons correspond to the difference between them in log odds of success on any item. Distances on the line between any two items correspond to the difference between them in log odds of success of any person. IRT models with more parameters do not permit the same concise representation as Figure 1, but the two 'constraints' referred to above embodying the widely held beliefs of all stakeholders suggest that this Rasch-based representation is the best (perhaps only) way to characterize the standard-maintaining situation.

The ideal theoretical solution to the standard maintaining problem would be to use a formal statistical equating design (e.g. Kolen & Brennan, 2004). However, in our context the usual way of establishing linkages between the score scales using formal statistical equating designs is not possible (Bramley & Black, 2008). In the Rasch / IRT framework, such linkages are achieved by having either 'common persons' (who attempt some or all items from both of the tests to be linked); or 'common items' that appear on both tests. (e.g. Wolfe, 2000;). Given such a linked design, there are two widely used ways of actually linking the scales. In the first, each complete test is calibrated separately, then the calibrations of the common elements (items or persons as per the design) are linked by calculating the constant shift needed to give the common elements

[3] A grade boundary (also known as a cut-score) is a point on the raw mark scale that separates examinees into two adjacent grade categories, e.g. 'A' and 'B'.

on one test the same average calibration as the common items on the other. Sometimes a multiplicative constant is also calculated in order to give the common elements on one test the same mean and standard deviation as those on the other. However, this raises some issues as to the substantive meaning of the unit (Humphry & Andrich, 2008). Arguably changes to the unit should not be made in the absence of a theoretical reason for doing so (for example, a theory that led to a prior hypothesis that the average discrimination of the common elements would be different in the two tests). The second common linking method is to calibrate all the data simultaneously by preparing a data set with 'missing data' for the responses to the non-common items or persons as illustrated in Figures 2a and 2b below. In this situation, the estimation algorithm implemented by the software handles the adjustments for scale origin (and unit) and automatically places all persons and items on a common scale, but with a potential loss of understanding and control for the analyst.

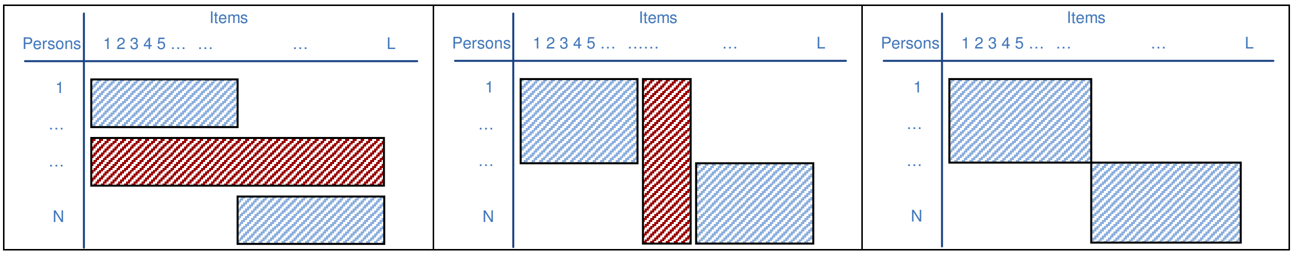

Figure 2a: Common person design 2b: Common item design 2c: No common elements.

The situation that we generally face is illustrated in Figure 2c, which illustrates a data set with no common elements. It is of course possible to make assumptions in order to link such data – for example that the average ability of the examinees on each test is the same, or that the average difficulty of the items is the same. However, as noted above, the second assumption would not be readily accepted by the stakeholders. The first assumption (usually made in terms of expected pass rates) is more acceptable when there is a large stable cohort of examinees and no particular reason to believe that they have changed much from one equivalent session to another – but to make this assumption introduces some circularity into the process, as described later in this paper.

The research that my colleagues and I have been pursuing for several years is directed at discovering the extent to which it is possible to use the judgment of experts to make the required 'common link'. If the standard Rasch model for dichotomous items is considered to be a two-facet model, with persons and items being the two facets, our approach can be considered as a one-facet model, using judges to make comparisons of a single facet using a paired comparison or rank-ordering approach.

The method has its conceptual basis in Thurstone's work on measuring subjective attributes (Thurstone, 1927a; 1927b). The main idea is that (a function of) the relative frequency of occasions on which one object is judged to be better than another (or, in more neutral terms, to possess more of the attribute than another) is an estimate of their separation on the scale to be created by the analysis.

Thurstone's original model used the normal distribution, but we have used the more tractable logistic formulation of his simplest Case 5 model, which is equivalent to the Bradley-Terry model for paired comparisons (Bradley & Terry, 1952). The connections between the Thurstone Case 5 model and the logistic model are discussed in Andrich (1978). The underlying theory and practical details of the method as we have applied it are explained in detail in Bramley (2005), Bramley (2007), and Bramley & Black (2008). The software we use to estimate the parameters is FACETS (Linacre, 2005).

Judgment about scripts

Initially, our investigations focussed solely on judgments of the quality of work produced by examinees (we refer to these pieces of work as 'scripts', but of course in principle the method could involve judgments of anything – e.g. artwork, videos of performance etc.). In this situation the task we ask the judges to carry out is to place packs of scripts (containing scripts from both tests) into rank order according to their perception of the quality of work produced, taking account of their perception of any differences in difficulty between the pair of tests involved. The mark totals (i.e. item and test total scores) are removed from the scripts so that they do not influence the judgments.

In order to compare the raw mark scales, scripts from across the effective mark range of both tests must be used, and arranged into packs in such a way that every script in the exercise is compared, directly or indirectly, with every other script. A direct comparison is obtained when scripts A and B appear in the same pack. An indirect comparison is obtained when script A and B do not appear in the same pack as each other, but both appear in the same pack as a third script, C. Indirect comparisons can also be much more indirect than this, when A and B are only linked via a series of intermediate scripts. The linking is necessary in order to derive a 'measure' of perceived relative quality for each script in the exercise. The relatively large number of scripts involved and the relatively small number of judges (and the usual practical constraints on cost and time for the judges) mean that it is not feasible for every judge to make every possible paired comparison. Therefore the information about total score is used as 'prior knowledge' in order to allocate scripts to packs in a way likely to maximise the information gained from each comparison. In other words, we avoid placing very good scripts in the same pack as very poor scripts because their relative ranking would be a foregone conclusion.

At the end of the exercise, after the ranking judgments have been analysed with a Rasch model[4], each script has a mark (raw score) and a 'measure'. The marks are on the same scale only within each test, but the measures are on the same scale across both tests. Therefore plotting mark against measure for each test separately allows the mark scales to be linked via the measures, as shown in Figure 3 on the next page.

[4] We have usually converted the rankings to paired comparisons (as suggested by Thurstone, 1931), although this does violate the requirement for local independence (Bramley, 2005; Linacre, 2006). However, experience has shown that using the Rasch Partial Credit Model (Masters, 1982) instead tends to give very similar outcomes, the main difference being that scale separation reliability is artificially inflated when converting rankings to paired comparisons.

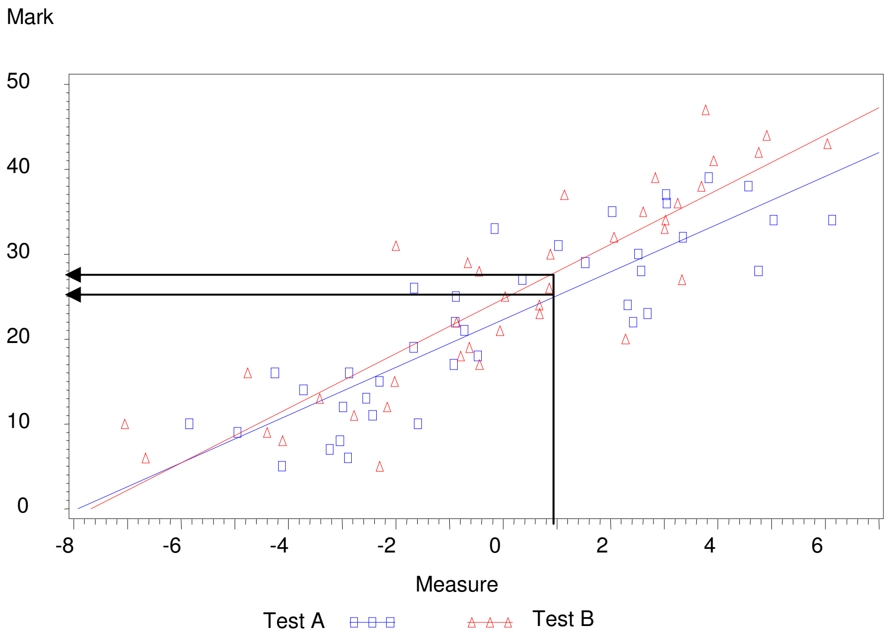

Figure 3: Example of typical outcome of a rank-ordering exercise with experts making relative judgments about script quality. A score of 25 on Test A corresponds to a score of 28 on Test B.

One clear indication of the confidence that one might have in the outcome is the correlation between the mark and the measure. If there is no relationship, or only a weak relationship, it is hard to justify the claim that the experts' judgments and the obtained marks reflected the same latent trait. Our practical experience has suggested that if the correlation drops below about 0.8 then the linking outcome is not robust enough. In looking for causes of low correlation, we have noticed that often it is associated with a low value for the scale separation reliability of the judgments, implying that measurement error could be responsible. In some cases this could perhaps be remedied by increasing the size of the study (more judges, more judgments per judge), but in others it may simply reflect the fact that no coherent 'common construct' was perceived by the judges.

We have argued that the outcome produced by the method cannot easily be validated in terms of whether it is 'correct', because the correct outcome, even if it can be sensibly defined, is not known in the circumstances where we use the method. We have therefore tried to validate the method by examining how robust the outcomes are to disturbances in various incidental aspects of a typical rank ordering study, such as the number of judges, the number of scripts per pack, and the method of summarising the mark-measure relationship. These investigations are reported in detail in Bramley & Gill (in press).

Judgment about items

More recently we have investigated the possibility of getting experts to make judgments about the difficulty of items, rather than about the quality of scripts. Certainly for multiple-choice tests it does not seem to make much sense to ask experts to compare strings of responses (either scored or un-scored) in terms of quality. And arguably making judgments about relative difficulty, if it can be done, addresses the issue of standard maintaining more directly than anything else. That is, it might be said that the only good reason to change a grade boundary or cut-score from one session of an examination to another is if there is evidence that the overall difficulty of the questions has changed (see 'ideal' reasoning below). Unfortunately, it is very difficult to disentangle question difficulty from examinee ability. In the Rasch model and other IRT models difficulty and ability are conceived conjointly, as 'two sides of the same coin'. In England, in the absence of pre-testing, when setting grade boundaries in GCE or GCSE examinations the difficulty of the questions tends to be judged by how well the examinees have scored on them, which introduces an unwelcome circularity into the process (see 'actual' reasoning below).

Ideal (example) chain of reasoning:

1. The questions are slightly easier than they were last session;
2. Therefore we should raise the grade boundary by x marks to compensate.

Actual (example) chain of reasoning:

1. The cohort of examinees is of slightly lower ability than it was at the last equivalent session, according to our information about prior attainment;
2. Therefore we expect a slightly lower cumulative percentage of examinees to achieve the grade;
3. If we raise the grade boundary by x marks the grade will be achieved by a slightly lower cumulative percentage of examinees than at the last equivalent session;
4. Therefore the questions must have been slightly easier than they were at the last equivalent session.

The difficulty with the 'ideal' scenario above is in establishing evidence for point 1. If there was some way to estimate the relative difficulty of the questions independently of how examinees scored on them, this would presumably be a good thing. If experts could accurately judge relative overall difficulty of examination questions, the grade boundaries could even be set before the examination is taken.

Of course, asking experts to make judgments about item difficulty is nothing new, as the large literature on standard setting methods (e.g. Cizek, 2001) testifies. In particular, two item-based standard setting methods have received a lot of attention – the Angoff method (Angoff, 1971) and the Bookmark method (see for example Karantonis & Sireci, 2006). To give a very brief summary of these methods – in the former, the experts make judgments about the probability that a 'minimally competent' examinee would succeed on each item. The cut-score is obtained by summing these judged probabilities over items and averaging across judges. In the latter, the items are presented in a physical booklet in increasing order of calibrated IRT difficulty and the judges place a 'bookmark' at the point where they think the minimally competent examinee would have a less than 66% probability of success. The cut-score is the raw score on the Test Characteristic Curve (see below) corresponding to the location on the latent trait of the 'bookmark' item, again averaged across judges.

Both the Angoff method and the Bookmark method are standard *setting* methods. They are supposed to provide a defensible means of arriving at an appropriate cut-score – that is, to answer the question 'how good is good enough'? However, in the context in which we are operating, most of the time the task is one of standard *maintaining* – finding the cut-score on one test that represents an equivalent standard to one set on a previous test. Therefore we would argue that a method that explicitly involves a comparative element is more appropriate than (for example) repeatedly applying a standard setting method to each new version of the test.

The rank ordering / paired comparison approach can be applied in a very straightforward way to judgments of item difficulty. Instead of packs of scripts containing examples from each test to be linked, now the packs contain items from each test. The task for the expert judges is simply to place the items into order of difficulty. There would seem to be at least two big advantages to this: i) the judges do not need to estimate any probabilities; and ii) they do not need to conceptualise a 'minimally competent' examinee. These two features are involved in both the Angoff and the Bookmark method, and are difficult to justify (see for example Impara & Plake, 1998; Boursicot & Roberts, 2006).

As before, it is necessary to make sure when designing the rank-ordering study that each item on each test is compared either directly or indirectly with every other item, in order that the perceived relative difficulty of every item can be estimated within the same frame of reference.

The main difference compared to the ranking of scripts arises once the relative difficulties have been estimated. Instead of plotting observed total score against estimated ability (perceived quality), the *perceived* relative item difficulties are treated as if they were *empirical* relative difficulties that had been obtained (for example) through a common item equating study. Thus it is possible to derive an expected score for a test comprising the items from each test in just the same way as if the tests had been constructed from a calibrated item bank.

The expected score on the test is the sum of the expected scores on each item for a candidate of a given ability. For a dichotomous item, the expected score is the probability of success (P) of person *n* on item *i*, as given by the equation for the Rasch model:

$$ln\ [P_{ni} / (1\text{-}P_{ni})] = \beta_n - \delta_i, \qquad (1)$$

where β_n represents the ability of examinee *n* and δ_i represents the difficulty of item *i*.

The expected score on the test is the sum of these probabilities across the *L* items on the test:

$$T_j = \sum_{i=1}^{L} P_i(\beta_j) \qquad (2)$$

where T_j is the expected score for examinees with ability level β_j, *i* denotes an item and $P_i(\beta_j)$ is obtained via equation (1).

The Test Characteristic Curve (TCC) is a plot of expected score on test against ability. The cut-score on the new test corresponding to the same ability as the known cut-score on the previous test can be determined, as in Figure 4 below.

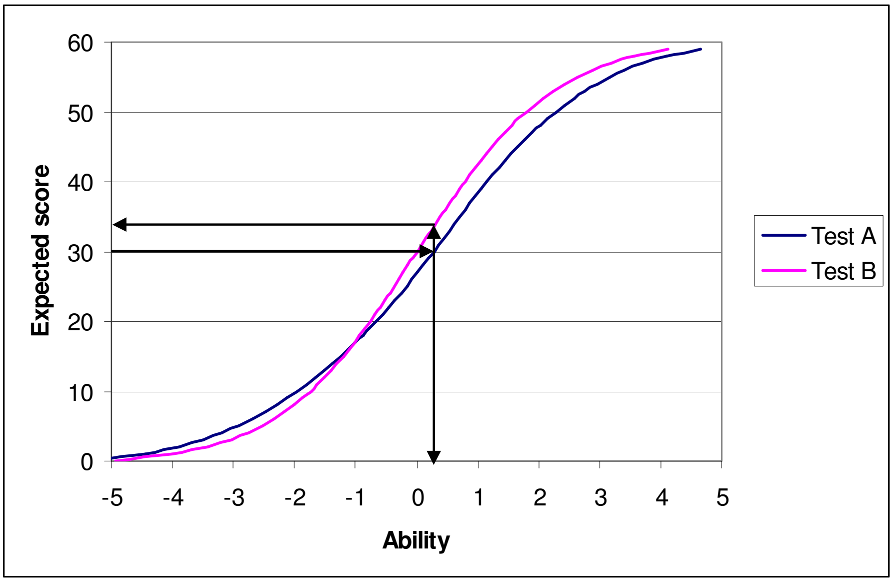

Figure 4: Example of typical outcome of a rank-ordering exercise with experts making relative judgments about item difficulty. A score of 30 on Test A corresponds to a score of 34 on Test B.

However, a plot like the one in Figure 4 provides no information about how well the expert judgments of relative difficulty corresponded to empirical information about relative difficulty, as indicated by facility values, or (if available) calibrated IRT difficulty estimates.

In the work we have done so far (Curcin, Black & Bramley, 2009; Curcin, Black & Bramley in prep), the correlations of perceived relative difficulty with empirical relative difficulty have been surprisingly variable and in general disappointingly low (although there have been some exceptions). Figure 5 below is an example. The correlation[5] for the November test was -0.46 and for the April test it was -0.34. Perhaps some non-linearity arising from ceiling effects on the facility values might have depressed the correlation, but it is clear that there was much less agreement between the judges' perceptions of relative difficulty and the empirical estimates than we have generally observed when experts are making judgments about relative script quality.

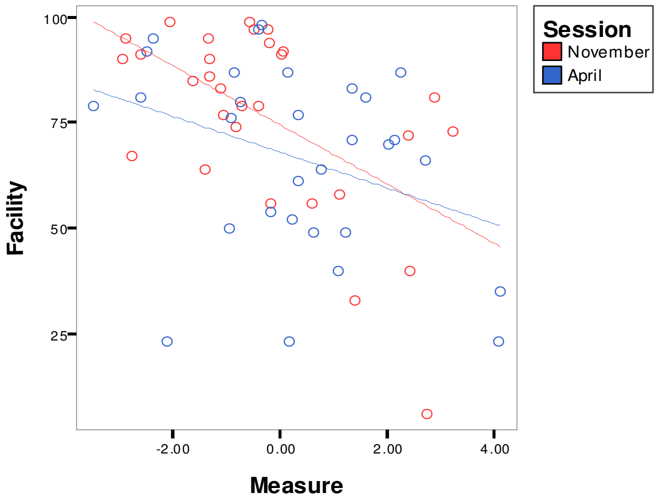

Figure 5: Example relationship between empirical facility value and estimated relative difficulty (from Curcin et al. 2009)

A more encouraging finding from the work on judgments of relative item difficulty has been that the estimates of relative difficulty do seem to be relatively reliable, in the sense of repeatable from one study to another. Curcin et al. (2009) found correlations of the order of 0.65 between estimates of perceived relative difficulty of the same items in studies carried out three months apart, and in Curcin et al (in prep.) the correlations were even higher (up to around 0.90). This suggests that the judges are sharing a perception of a 'common construct' of item difficulty. Unfortunately (at the moment) we cannot say with much confidence that this is the same latent trait that underlies examinee performance on the items. We are currently exploring whether judges who have had experience of *writing* items can be distinguished from judges who are subject matter experts but with no experience of writing items; and we are collecting 'think-aloud verbal protocols' from judges while they are carrying out the ranking task to see what features of the questions influence their judgments, and more generally what thoughts seem to be going through their minds while they do the task. We are also asking the examinees themselves to rank-order items in terms of perceived difficulty to see if they have a better idea than the experts of how difficult people like them will find the items!

Discussion

There are several ways of estimating the parameters of the Rasch model, but one that is of particular interest in this context is the 'pairwise' algorithm (Choppin, 1968; Wright & Masters, 1982). Considering any pair of items A and B, if N_{10} people have got A correct and B incorrect, and N_{01} people have got B correct and A incorrect, then $\ln (N_{01} / N_{10})$ is an estimate of $\delta_A - \delta_B$ where δ_A and δ_B are the locations of items A and B on the trait. This can be viewed as treating the examinees as 'judges' of which of item A and B is the more difficult. There is thus a close connection between the 'one-facet' expert judgments of relative item difficulty and the more

[5] We would hope that this correlation was high and negative, because higher facility corresponds to lower difficulty.

usual ‘two-facet’ analysis of person-item response data – providing of course that the same latent trait underlies both sets of data. Essentially we are asking the judges to rank-order items in terms of δ. It is possible (even likely) that the scale produced by the analysis of the judges’ rankings will have a different unit size from the scale produced by analysis of response data. However, there is no reason to suppose that any difference in unit size (perceived difficulty scale compared with empirical difficulty scale) would be different depending on which test the items had come from, so the fact that we are comparing ‘like with like’ when linking the two tests via the TCCs (see Figure 4) should mean that the unit size issue is not so relevant to this method of standard maintaining.

For judgments of scripts, on the other hand, the analogy with the pairwise algorithm does not work. This is because the thing being judged (the script) consists of all the items in the test plus the examinee’s responses to them. Also, half of the scripts in each pack will come from different tests, so the judges have to make some allowance for any perceived differences in difficulty between the two tests. Essentially we are asking the judges to rank-order examinees in terms of β while presenting them with (un-scored) evidence of the outcome of (β-δ_i) for each item *i* on the test. The judgment is therefore (at least on the face of it) much more complex than the judgment we ask the judges to make when rank-ordering items in terms of difficulty. They have to allow for the different perceived difficulty values δ_i arising from the different tests, and aggregate their judgments across items.

This raises the question of why the observed correlation of empirical difficulty (e.g. facility value) with perceived difficulty has tended to be *lower* than the observed correlation of empirical ability (test score) with perceived ability. If the latter judgment is more complex than the former, and even requires the former as a component part, we might expect it to be more vulnerable to both systematic and random error, and hence to produce lower correlations. Perhaps the answer lies in the way the judgments about scripts are aggregations of judgments about performance on items – if most error in these judgments is random then the aggregation will lead to cancelling of error. Using a further analogy, this time with adaptive testing, perhaps as the judge reads each script they are mentally adjusting an estimate of β_n upwards or downwards as they encounter good or poor answers respectively. By the time they reach the end of the script they have arrived at a relatively stable (and accurate) estimate of β_n, even if their estimates of the individual δ_is of the items were not particularly accurate[6]. A further reason could simply be one of familiarity – experts are more used to making judgments of the quality of work when they mark (score) the scripts using the mark scheme (scoring rubric). Making judgments of relative item difficulty is likely to be a less familiar task.

Is the latent trait underlying the judgments of ‘perceived quality’ of scripts the same as the latent trait underlying the performance data? This is not an easy question to answer, but we have approached it by trying to understand what features of scripts influence the experts’ perceptions of relative quality. Bramley (2009) reports on a study where the effect of experimentally manipulating four different features of scripts was investigated. Interestingly, the results showed that the judges were sensitive to changes to some features of scripts that did not affect the total score, such as replacing some incorrect responses with blank responses, or changing the profile of correct and incorrect responses within the script. The effects observed were quite small, but the results do cast some doubt on the idea that the two latent traits are identical. Future experimental work could carry out an equivalent investigation for perceptions of relative item difficulty, by systematically manipulating some features of the items involved in a rank-ordering study.

In conclusion, we have promoted paired comparison and rank-ordering methods as tools for using the judgment of experts to maintain the standards represented by cut-scores on

[6] Just to be absolutely clear – I am not suggesting that the judges are consciously or unconsciously actually estimating any numerical parameters when they make judgments about items or scripts! This is merely a way of thinking about what they are doing using the Thurstone/Rasch conceptual framework.

educational tests in situations where conventional methods of statistical equating are unavailable or inappropriate. The method can be applied to both judgments about scripts and judgments about items. The task required of the judges is easy to explain to them and does not require judgments of probabilities, conceptualisations of hypothetical examinees, or comparisons with arbitrary or vague criteria. Defensible decisions in high-stakes testing contexts require relevant evidence and rationales for determining how much weight that evidence should be given. We have recommended (Black & Bramley, 2008) that evidence from paired comparison and rank-ordering studies should be treated as just one source of evidence among several when making decisions on where to set cut-scores in order to maintain standards in UK high-stakes school examinations like the GCSE and GCE A level. The Thurstone and Rasch approaches to measurement of subjective attributes provide a conceptual framework in which the issues involved in standard maintaining can be brought into a clearer focus.

References

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42(1)*, 69-81.

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

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement*. (pp. 508-600). Washington DC: American Council on Education.

Black, B., & Bramley, T. (2008). Investigating a judgmental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education, 23(3)*, 357-373.

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Borsboom, D. & Zand Scholten, A. (2008). The Rasch model and conjoint measurement theory from the perspective of psychometrics. *Theory & Psychology, 18(1)*, 111-117.

Boursicot, K., & Roberts, T. (2006). Setting standards in a professional higher education course: defining the concept of the minimally competent student in performance-based assessment at the level of graduation from medical school. *Higher Education Quarterly, 60(1)*, 74-90.

Bradley, R.A., & Terry, M.E. (1952). The rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrica, 39*, 324-345.

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

Bramley, T. (2007). Paired comparison methods. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards.* (pp. 246-294). London: Qualifications and Curriculum Authority.

Bramley, T. (2009). *The effect of manipulating features of examinees' scripts on their perceived quality.* Paper presented at the Association for Educational Assessment – Europe (AEA-Europe) annual conference, Malta, November 2009.

Bramley, T. & Black, B. (2008). *Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work.* Paper presented at the Third International Rasch Measurement conference, University of Western Australia, Perth, January 2008.

Bramley, T. & Gill, T. (in press). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education.*

Choppin, B. (1968). Item bank using sample-free calibration. *Nature, 219*, 870-872.

Cizek, G.J. (2001). *Setting Performance Standards: Concepts, Methods and Perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Curcin, M., Black, B. & Bramley, T. (2009). *Standard maintaining by expert judgment on multiple-choice tests: a new use for the rank-ordering method.* Paper presented at the British Educational Research Association annual conference, University of Manchester, September 2009.

Curcin, M., Black, B. & Bramley, T. (in prep.) *Towards a suitable method for standard maintaining in multiple-choice tests: capturing expert judgment of test difficulty through rank-ordering.* Paper to be presented at the AEA-Europe annual conference, Oslo, Norway, November 2010.

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.

Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal, 5(2)*, 211-220.

Humphry, S., & Andrich, D. (2008). Understanding the unit in the Rasch model. *Journal of Applied Measurement, 9(3)*, 249-264.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35(1)*, 69-81.

Karantonis, A., & Sireci, S.G. (2006). The bookmark standard-setting method: a literature review. *Educational Measurement: Issues and Practice, 25(1)*, 4-12.

Kolen, M.J., & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices.* (2nd ed.). New York: Springer.

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology, 18(1)*, 89-109.

Linacre, J.M. (2005). FACETS Rasch measurement computer program. www.winsteps.com

Linacre, J.M. (2006). Rasch analysis of rank-ordered data. *Journal of Applied Measurement, 7(1)*, 129-139.

Maraun, M.D. (1998). Measurement as a normative practice: implications of Wittgenstein's philosophy for measurement in psychology. *Theory & Psychology, 8(4)*, 435-461.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47(2)*, 149-174.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 335-383.

Michell, J. (2008). Conjoint measurement and the Rasch paradox. *Theory & Psychology, 18(1)*, 119-124.

Thurstone, L.L. (1927a). Psychophysical analysis. *American Journal of Psychology, 38*, 368-389.

Thurstone, L.L. (1927b). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Thurstone, L.L. (1931). Rank order as a psychophysical method. *Journal of Experimental Psychology, 14*, 187-201.

Wolfe, E.W. (2000). Equating and Item Banking with the Rasch model. *Journal of Applied Measurement, 1(4)*, 409-434.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14(2)*, 97-116.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16(4)*, 33-45,52.

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, NJ: Lawrence Erlbaum Associates.

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