Psychometric characteristics of integrated multi-specialty examinations: Ebel ratings and unidimensionality

Matt Homer[1], Jonathan Darling and Godfrey Pell

Abstract

Over recent years UK medical schools have moved to more integrated summative examinations. This paper analyses data from the written assessment of undergraduate medical students to investigate two key psychometric aspects of this type of high stakes assessment.

Firstly, the strength of the relationship between examiner predictions of item performance (as required under the Ebel standard setting method employed) and actual item performance ('facility') in the examination is explored. It is found that there is a systematic pattern of difference these two measures, with examiners tending to under-estimate the difficulty of items classified as relatively easy, and over-estimating that of items classified harder. The implications of these differences for standard setting are considered.

Secondly, the integration of the assessment raises the question as to whether the student total score in the exam can provide a single meaningful measure of student performance across a broad range of medical specialties. Therefore Rasch

[1] Corresponding author. Leeds Institute of Medical Education, School of Medicine, University of Leeds, Leeds LS2 9JT

measurement theory is employed to evaluate psychometric characteristics of the examination, including its dimensionality. Once adjustment is made for item interdependency, the examination is shown to be unidimensional with fit to the Rasch model implying that a single underlying trait, clinical knowledge, is being measured.

Bibliographical details

Dr Matthew Homer is a Research Fellow at the University of Leeds working in the both the Schools of Medicine and Education. His work is focussed on the quantitative elements of a range of research projects, and he provides general statistical support to colleagues. His research interests include aspects of psychometrics and the statistical analysis of large datasets.

Dr Jonathan Darling is Senior Lecturer in Paediatrics and Child Health in the School of Medicine at the University of Leeds. He has a particular interest in undergraduate medical education, especially all aspects of assessment. He chairs the MBChB Year 4 course management team, and is lead for Year 4 written assessments.

Godfrey Pell is a senior statistician who has a strong background in management. Before joining the University of Leeds he was with the Centre for Higher Education Practice at the Open University. His current research interests centre on standard setting for practical assessments in medical undergraduate programmes.

Introduction

The use of integrated summative examinations in the later years of medical schools has grown over recent years, particularly in the UK. These examinations aim to assess students' clinical knowledge across a specific range of medical specialties whereas previously these specialisms were assessed in separate exams. This paper analyses data from one such examination to investigate two key separate psychometric aspects of this type of high stakes assessment:

(i) The extent to which the standard setting is accurate in the sense that examiner predictions of item performance used to set the pass/fail cut-score reflect actual item performance in the exam; and,

(ii) Whether or not student performance in the exam can be adequately summarised in a single mark, and hence whether pass/fail decisions can be accurately made using this single outcome.

Before further discussing these two aspects of the examination , the paper begins with a description of the structure of the examination and of the standard setting methodology that is employed.

The format of the examination

At the end of the fourth year of a five year undergraduate medical programme at the University of Leeds in 2008, the cohort of 244 students were given an integrated summative written examination in three parts as follows[2]:

[2] The other strand of the summative assessment is a practical test of clinical competence and skill.

TABLE 1 HERE

Extended matching questions (EMQs) as employed at Leeds are grouped into sets of five clinically-related questions all presented with the same theme and option set, where a short vignette for each question describes various details such as, for example, a patient's symptoms (Case and Swanson 1998, Chapter 6). The student's task in each case is to select the most appropriate option from the list, for example this could be the most likely diagnosis, chosen from list of potential diagnoses

For the slide questions, the students are presented with slides in a PowerPoint presentation which they scroll through at will. These show, for example, pictures of a patient with a certain condition, or the results of a medical procedure such as an x-ray. Students have an average of 75 seconds per slide, and have to answer either one or two associated single best answer questions for each slide. Each slide is worth two points, so if a slide has only one item, this item is worth two points, but if it has two items then each is worth 1 point.

In the EMQ papers a 'question set' refers to a group of questions presented with one theme and option set, whilst in the Slide paper it refers to all questions associated with a particular slide.

All of the questions in the examination are written by medical educators at Leeds or elsewhere. A small proportion of the 330 items are re-used from previous years but all others are written specifically for each particular examination.

The standard setting methodology employed

The overall pass mark for the annual examination is set using a variation on a modified Ebel approach (Ebel 1972, Skakun & Kling 1980, Cizek & Bunch 2007, Chapter 6). The essence of the Ebel method is that examiners must hold in their minds an idea of the typical performance of students who are at the borderline between acceptable and unacceptable levels of performance. The annual standard setting process has two distinct aspects as follows:

Item difficulty and relevance categorisation

A group of between three and five number examiners categorise each individual question according to two separate dimensions, with three categories in each: 'difficulty' for the borderline candidate in the opinion of the examiner (Easy, Moderate or Difficult) and 'relevance' in terms of the aims of the curriculum (Essential, Important, Additional) (Cizek & Bunch 2007, 76-90). The examiners are drawn from the range of specialities represented within the exam, and rate questions from all specialities, not just their own. As described in the literature, the Ebel method requires a discussion amongst the examiners in order to come to an agreement on the categorisation of items (Cizek & Bunch 2007, 76). However, calling a group of senior medical professionals together for such a meeting each year often proves difficult in practice due to logistical and time constraints. Therefore although some question-rating is done by discussion in face-to-face meetings, most is done independently and then individual examiner judgements are averaged and rounded to the nearest integer to produce a single difficulty and relevance rating for each of the 330 questions.

Table 2 summarises the distribution of averaged examiner ratings of the 330 items making up the examination. The majority of items (55.8%) are classified in the middle category for both difficulty and relevance. Two cells contain no items (essential-difficult, and easy-additional).

TABLE 2 HERE

Expected borderline candidate performance

For each of the nine possible combinations of ‘difficulty’ and ‘relevance’, the examiners estimate the proportion of borderline candidates who would be expected to know the answer to that particular type of question (Cizek & Bunch 2007, 77). These percentages are determined through consensus, and are shown in Table 3. Hence, for example, in the examiners’ judgement, 45 per cent of borderline students would be expected to know the answer to a question that was tagged as *Moderate* but *Important*.

TABLE 3 HERE

The values in this table are likely to vary from institution to institution as they will each have their own shared understanding of the precise meaning of the relevance and difficulty categories for their students according to their own curricular aims (Cizek & Bunch, 2007, Chapter 5). There is evidence in the medical education literature of varying standards across medical schools (Boursicot et al 2006).

Once the two processes are complete, the single pass mark for the examination is calculated in two stages as follows:

1. Using the item categorisation in Table 2 and the expected performance matrix in Table 3 , the probability of a correct response for each item is automatically calculated in a spreadsheet containing all the ratings, and using an adjustment for guessing depending upon the number of available options available. For example, if there are six options for an item, with an expected borderline percentage of 75%, the adjusted value would be 75% + 25%/6 = 79.17%[3] based on the assumption that one sixth of those students who did not know the answer would correctly guess. This final adjusted value can be thought of as the predicted probability that a borderline candidate will get the item correct as judged by the group of examiners This value is referred to throughout the paper as the predicted facility for the item.
2. The total expected mark for each of the three papers is then added up using a weighting of 2:2:1 for papers 1, 2 and 3 respectively. This gives the overall pass mark as set by the Ebel process. In other words, students who achieve this mark or above pass the written examination, and those below it are deemed to have failed.

Psychometric issues with the written examination

In high stakes assessment, one should constantly strive to monitor and improve the general quality of any examination in terms of, for example, its reliability and validity

[3] More generally, the correction employed is given by (100-e)/n where e=expected percentage of borderline students knowing the correct answer, and n=number of options for the item.

(Streiner & Norman 2003, chapters 8 and 10 respectively). In the current context, there are two distinct, but related, key psychometric strands to this investigation:

- the Ebel standard setting methodology employed – are the examiner estimates of item performance (i.e. the predicted facilities) sufficiently accurate to ensure that standards are appropriate and are maintained over time?
- the items themselves that make up the assessment – do they form a sufficiently unidimensional measure of performance so that a single score carries sufficient information for making pass/fail decisions?

These two aspects of the assessment are considered separately in the analysis. Later, in the Discussion, they are considered together.

Standard setting using the modified Ebel method

There are advantages and disadvantages to whichever standard setting procedure is employed, and there are bodies of published arguments for and against each (Norcini 2003). With regard specifically to Ebel-based methods, there are studies that compare Ebel standard setting with other such methods (Downing et al 2003, Skakun & Kling 1980). There is also a recent study that investigates the impact of the number of examiners, and the extent of their deliberations, on the reliability of pass marks derived from Ebel judgments (Fowell et al. 2006). However, there is little evidence of previous research into precisely how well examiner estimates of item difficulty under Ebel correspond with actual item performance.

There is an element of circularity in the identification of the borderline group of students. They are identifiable in a post hoc analysis solely on the basis of being ‘close’ to the pass mark (within, say, 1 or 2 standard errors of measurement (SEM) of it (Streiner & Norman 2003, 142), but this standard has been set using precisely those judgements under scrutiny in this study. This problem is presumably one reason why there has been little research into the predictive accuracy of the Ebel examiner judgements of item difficulty.

The Ebel process asks examiners to make item-level judgements before the exam about the standard that **should** be achieved by the minimally competent candidate. Hence, it is primarily a test-centred standard-setting process (Cizek & Bunch, 2007, p9), where judgments are focussed on the items making up the examination rather than on the examinees themselves. Examiners have to envisage a borderline candidate and then decide how relevant and how difficult the question is for that (imaginary) person. In practice, these judgements are actually examiner predictions of the standard that **will** be achieved by the borderline candidate, and therefore a strong relationship between examiner-rating and candidate/item performance would be expected. If there is evidence that this relationship is not especially strong then this undermine the Ebel standard setting methodology, particularly in terms of the maintenance of comparable standards over time.

Our first analyses were done to estimate the strength of association between item difficulties as derived from examiner judgements under the Ebel method, and item difficulties derived from classical test theory as the percent of the cohort answering each correctly (Crocker & Algina 1986, 311-2) .

The measurement of a single underlying construct

The summary of performance in the written assessment is a single mark, on the basis of which students either pass or fail. However, the examination is made up of two distinct assessment types (EMQ and Slide), and contains items across five different medical speciality areas. This complexity raises the key issue of whether or not this type of examination, with a single overall measurement outcome, is in fact measuring a single underlying construct (Hambleton 1991, Chapter 2) say, clinical knowledge. If this were shown to not be the case, then a single mark might not contain enough information to provide an adequate summary of performance, and pass/fail and other grading judgements might need to be made on the basis of a more complex decision-making process, perhaps involving sub-sections of the examination that are themselves demonstrably unidimensional.

To investigate this issue, our second analyses used Rasch-based statistical techniques (Rasch 1960/1980, Bond & Fox 2007) to study a range of psychometric properties of the examination, including the central question of unidimensionality, but also considering other important issues such as individual item fit, internal consistency reliability, and the overall targeting of the exam.

Historically, there has been some criticism, particularly in the UK, of the Rasch model (Goldstein 1979), but internationally the technique has been employed for many years in high-stakes assessment, both in and outside medical education. For a review of these issues and for an introduction to the key elements of Rasch modelling see Panayides et al. (2009), who argue that some of the earlier criticisms of the Rasch model were based on fundamental misunderstandings of the method.

Standard setting and Rasch modelling

There is evidence in the literature of attempts to facilitate standard setting using modern measurement theory approaches (such as Rasch modelling), particularly under Angoff-based methods (Maccann 2009, O'Neill et al. 2005). Angoff standard-setting methods (Cizek & Bunch, 2007, chapter 6) are similar to those under Ebel, except that the examiners consider each individual item separately in detail, and judge the likely performance of the minimally competent student when presented with it. In essence, there is no equivalent of Ebel's relevancy/difficulty grid, Table 2, but rather just a list of expected performances, one for each item. However, these previous studies have tended to use Rasch-based estimates to **inform** examiner judgements rather than to consider directly, as this paper will, the psychometric properties of the exam from a Rasch perspective.

Methods

This paper has two main overall objectives and these will be investigated in turn;

1. To consider the accuracy of examiner predictions of item performance under the Ebel method.
2. To investigate the extent of the unidimensionality of the assessment using Rasch measurement theory.

1. Standard setting using the modified Ebel method

The borderline group of students

Over the last five years, the borderline group for this examination is always in the 10th decile based on total score – there are usually approximately 10% of students awarded the lowest of four pass grades, with some of these clear passes rather than being actually borderline based on the SEM criterion. There are also an additional 1 or 2 % annually who fail. Hence, for this study, when attempting to focus on the borderline group as per the Ebel examiners, the 10th decile students are used as a proxy for this group, always bearing in mind the limitation that it contains a proportion of students who are not actually borderline according to the Ebel definition. For comparison, where appropriate, additional results for the other deciles are also presented.

Correlation analysis

The aim of these analyses was to estimate the strength of association between Ebel-derived item difficulty estimates (as described in the introduction) and difficulties derived as the percent of the current cohort answering each correctly (Crocker & Algina 1986, pp. 311-2). This latter measure is usually referred to as the item facility.

Pearson correlations are used to compare the facility for each of the 330 items in the examination (i.e. the percentage of students answering correctly) with the corresponding Ebel item-level ‘facilities’ as judged in advance by the examiners. The correlation analysis is done both for the student group overall (n=244), but also within student decile (based on the overall performance in the exam) to control for

overall ability. The analysis is also carried out separately across the three constituent papers making up the exam to allow for differences across these to emerge.

SPSS version 15.0 was used for all statistical analysis in this paper.

Bland Altman plot

Correlations only measure the strength of the linear relationship between two variables, rather than the extent of their actual agreement. As an alternative, graphical methods (Bland-Altman plot, Bland & Altman 1986) are also used to gain additional insight into the extent of agreement between these two 'measures' of item difficulty. In order to focus the analysis on the key borderline group, the difference between the Ebel predicted facility and the actual item facility for the 10th decile students only is plotted (on the y-axis), against the actual item facility on the x-axis. In a standard Bland-Altman plot, the x-axis is used to plot the mean of the two measures, based on the principle that the true measure is not known, and that therefore the mean provides the best estimate of this true (unknown) value. In this study, it is assumed that the actual item facilities provide the true measures of item difficulty.

2. The measurement of a single underlying construct

Rasch modelling is a statistical technique for analysing item responses (Bond & Fox 2007). The theory is based on the underlying assumption that the probability of responding correctly to a dichotomous item is a (logistic) function only of the difference between the student's ability, and the item's difficulty (Bond & Fox 2007,

Appendix A). Rasch analysis provides a parsimonious and rigorous theory of measurement that allows detailed scrutiny of scored assessments.

The Rasch model is testable, and in order for the resulting person ability and item difficulty estimates to be correctly employed there are key properties of the data (i.e. of the pattern of responses) that must hold, at least approximately. These include unidimensionality (i.e. that the exam is measuring a single underlying construct), local independence (that once ability has been accounted for, item responses are independent), and invariance (that item difficulty estimates and person ability estimates are constant across suitable samples) (Bond & Fox 2007, 32-4, 172 and 69-99 respectively). Using Rasch software, such as RUMM2020 (Andrich et al. 2002), these three key properties of the assessment as a whole are testable, and the extent to which they hold in our data will be investigated.

The objectives regarding the Rasch analysis are as follows:

- to investigate the extent to which the exam is well-targeted to its students.
- to quantify the fit to the Rasch model of the exam; and
- to assess the degree of unidimensionality and reliability of the exam;

Where necessary, inter-item dependencies at the question set and specialty level are taken account of through the grouping of related items into testlets (Wainer & Wang 2000, Wainer et al 2007) based on the common theme, and then specialty.

Results

1. Standard setting using the modified Ebel method

Correlation analysis

Figure 1 shows a graph of estimates of the Pearson correlations between the predicted Ebel item-level marks and the actual item facilities resulting from student performance in the examination. These were calculated for the exam overall, and for each of the three papers separately. The actual facilities (i.e. percentage of students answering the item correctly) were calculated for the whole group of examinees, and for each decile based on overall (non-weighted) total student score in the examination. This separation by decile was carried out to investigate whether or not the correlations varied according to overall student ability - one might have perhaps expected that there could have been a stronger correlation for those deciles that were closest to the borderline performance examiners were attempting to judge during their consideration of items (as already stated, this borderline group has always been located in the 10^{th} decile for this type of examination).

FIGURE 1 HERE

It is recognised that the analysis by decile might provide underestimates of the correlation coefficients due to the effect of restricting the range of scores in this way. However, as Figure 1 shows, the similarity between the magnitude of the overall correlation and the within-decile correlations suggests that any such effect is not particularly large for this data.

There is some variation in the size of the correlations by exam paper, with paper 2 showing stronger correlations generally (the standard error of the correlation estimates vary from 2.37% to 2.72%, so 95% confidence intervals for these have a half-width of the order of 5%). More importantly, the typical size of the correlation coefficient (R), whilst statistically different from zero in all but four (out of 44) cases ($p<0.005$, Bonferroni corrected), is not particularly large (of the order of 0.5 or less generally). This indicates that only around a quarter of the variation in the actual facilities of the items is shared with the examiner judgements of their relevance/difficulty (R^2).

Bland-Altman plot

It is known that correlations do not always provide a good indication of the degree of agreement between two different apparent measures of the same thing (Bland & Altman 1986). Even high correlations can be misleading in this regard. As an alternative, a variation on the Bland-Altman plot is used to assess the agreement between two apparent measures of the same thing. For each item in the exam, Figure 2 plots for students in the 10^{th} decile:

- Ebel 'predicted facility'(i.e. the Ebel-determined passing 'score' for the item) minus the actual item facility (on the y-axis); against ,
- the actual item facility (on the x-axis).

‘Actual item facility’ here (and elsewhere) is the proportion of the sample getting the item correct, so that high facilities correspond to easy items and vice versa.

A positive y-value in Figure 2 implies that the examiner predictions are an over-estimate of item facility and that therefore the item proved more difficult for those in the 10th decile than they had judged it to be.

The middle dashed horizontal line is at the mean difference between the two ‘facilities’ (10.6%), and the upper and lower lines indicate this mean ± 2 standard deviations of the difference. For strong agreement one would expect most y-values to be close to zero, but the plot shows that is an overall bias in the examiner judgements – they are under-estimating the item facility for this group of students by a mean of 10.6% (as shown in Figure 2 by the middle dashed horizontal line). Since the top of the 10th decile is very likely to contain some non-borderline students, this average over-estimate of item difficulty might be expected. The sample contains a proportion of students who are at a higher ability level than that which the examiners’ are intended to focus on when making the judgements.

Importantly, there is a downward trend across the plot implying that there is a systematic tendency amongst the examiners to under-estimate the difficulty of items classified as relatively easy, and over-estimate that of those classified harder. This is not what one would expect when confining the analysis to a group that, whilst including the borderline candidates, also includes a proportion of higher ability. Assuming accurate predictions, one would expect a consistent picture across the difficulty range (the x-axis) of negative differences. In addition to this systematic

component, for each value along the horizontal axis (i.e. actual facility), there is considerable vertical spread implying that there is secondary, more random component to the difference between the two measures, demonstrating that examiner judgements under Ebel also show a degree of additional error in comparison to actual item performance.

.

FIGURE 2 HERE

2. The Rasch analysis

The Rasch methodology and protocols employed throughout the analysis presented here are based on that of Tennant & Conaghan 2007. The software used in the analysis was RUMM2020 (Andrich et al. 2002).

Extreme items

There are two items that were 100 per cent correct, one each in papers 1 and 3, and one item in paper 1 that was 0 per cent correct. Such ‘extreme’ items have to be removed from the analysis since according to the Rasch model, the estimates of their difficulty will not be finite (Bond & Fox 2007, Appendix A). This leaves 327 items in the remainder of this analysis.

The ratio of the number of items to the relatively small number of persons (327 to 244) in the data might be considered a problem in terms of producing robust item

and person estimates. However, the key requirement is merely sufficient numbers of either items or persons in the data (Linacre 1994 states 30 or more). This is an important advantage of Rasch over 2- and 3-parameter IRT models, where the greater number of parameters being estimated implies the need for larger sample sizes, usually on a scale only existing in national assessments, and certainly much larger than those in the annual cohort of the typical medical school (for the 2-pararmeter logistic model, Champlain 2010 suggests 500 or more examinees, although there are many other estimates in the literature).

Individual person and item fit

There are five students (n=244, 2.0%) with significant misfit to the Rasch model expectations (fit residual greater in magnitude than ±2.5). Since it might be expected that approximately one per cent of students would show misfit at this level (in normally distributed data, 1.2% of the distribution is outside of Z=±2.5) this is not considered a problem in the analysis.

At the individual item level, there are 10 items (n=327, 3.1%, 2 in paper 1, 4 in paper 2 and 4 in paper 3; spread across specialties) with significant misfit, either in terms of chi-square probabilities (Bonferroni-corrected p-values less than 0.0001), or fit residuals (again>2.5). These are items that do not fit the expected pattern of performance under the Rasch model and should be prioritised for review by the item writers/assessors.

The overall targeting of the exam

It is important that an exam is appropriately targeted to the students taking it (Tennant & Conaghan 2007). Put simply, the exam as a whole should not be too easy or too hard or else the Rasch estimates of item difficulty or student ability will be poor (i.e. with large standard errors), and the reliability of the test will be low so that pass/fail decisions based on it will also be poor. Rasch parameter estimation places items and students on the same difficulty/ability scale, that is, it produces conjoint measurement of items and students (Bond & Fox 2007, 262-265). The 'location' on the Rasch scale represents its difficulty (item) or level of ability (student)..

The difficulty scale is always centred on zero (logits), representing the average item difficulty. The mean student location, calculated by the software as 1.323, standard deviation 0.489, indicates that the average student ability is higher than that of the average item difficulty. This is demonstrated graphically in the student-item location distribution (Figure 3), where the horizontal scale is (conjointly) ability/difficulty. It can be seen that students (upper graph) tend to be higher up the scale than do the items (lower). Overall, however, the figure shows that the examination was well-targeted to the students since there is broad overlap between the two distributions. It could, however, be argued that there were too many easy items (the long tail to the left of the lower graph), although there may be good pedagogical reasons for the inclusion of such items in the exam, for example, when it is important to test that students understand key elements of the curriculum.

FIGURE 3 HERE

Fit and unidimensionality of the exam as whole

Table 4 shows the results of three separate Rasch analyses, each carried out at a different level of focus in terms what exactly constitutes a scored 'item' in the analysis. As will be seen in the next three sub-sections, the unidimensionality and 'fit' of the exam is brought into question in the initial analysis, and it is only through grouping items by specialty into testlets (i.e. subsets of related items, Wainer & Wang 2000, Wainer et al 2007) that the failure to establish local independence can be overcome and unidimensionality established.

1. Analysis without grouping of items

The first analysis considers the exam as made up of 327 individual (ungrouped) items. For each item in the exam and each person taking the exam, Rumm2020 computes the difference between the observed performance and the Rasch model expectations of the performance. These are the item and person residuals respectively (Bond & Fox 2007, Chapter 12), and Table 4 (first row) shows that the mean residual item and person fit estimates are reasonable in this analysis (residual mean fits are expected to be 0, with standard deviation 1). There is, however, evidence of significant overall misfit to the Rasch model ($p<0.000001$) based on the item-trait interaction statistics which tests the extent to which persons of differing ability agree on the ordering of difficulty of the items. The internal consistency reliability of the examination, as measured by the Person separation index, which estimates how well persons are differentiated by the exam (Bond & Fox 2007, Appendix A), is good at 0.911.

The final column of Table 4 shows the results of a strict statistical test of unidimensionality. This uses the group of twelve items (formally, measurement points) with the *largest* loadings on the first principal component of the residuals to estimate person abilities (Tennant & Conaghan 2007). A second set of such estimates is also calculated using the group of twelve items with the *smallest* loadings on the first principal component of the residuals. If the exam is unidimensional then the person estimates from these two groups of items should not differ significantly. However, the results indicated that 10.66 per cent of person estimates are significantly different comparing these two sets of items, when one would expect only five per cent only by chance given unidimensionality - where the confidence interval for the estimated percent difference, shown in the final column in Table 4, includes the value 5% (row 1, items ungrouped), it is interpreted as evidence against unidimensionality).

The final aspect of the analysis of the ungrouped items involves investigating the extent of any response dependency in the data using residual correlations. Once the main Rasch factor, ability, has been extracted, the degree of correlation remaining amongst items should be small (Tennant & Conaghan 2007). If this is not the case this is evidence of a failure of local independence. Analysis using Rumm2020 shows that there is indeed some evidence of response dependency with, for example, 37 residual correlations greater than 0.3 (out of total number of 327×326/2=53,301 in total, 0.066%).

As a result of evidence of overall misfit to the Rasch model, and failure to demonstrate both unidimensionality and local independence, the items are systematically grouped together in two additional Rasch analyses in order to reduce local dependency and help mitigate problems with multidimensionality (Wainer & Wang 2000). These further analyses are described in the following two sub-sections.

2. Grouping items into testlets based on a common theme

In an EMQ-type exam, sets of items have a common theme and might naturally be expected to correlate with each other even after the main (ability/difficulty) factor has been accounted for. Hence, a second analysis was carried by grouping the items into 108 'testlets' based on each theme (resulting in 24 testlets in paper 1, 24 in paper 2 and 60 in paper 3; see Table 1). Note now that these testlets are polytomous rather than dichotomous (i.e. a person's score in not limited to just 0 or 1 as is the case for single items, but to 0 to 5 in the case of an EMQ with five questions to a set).

Table 4 (second row) shows the outcomes of this analysis, but the earlier problems of the first analysis in terms of non-fit to the model remained (overall chi-square test of misfit highly significant; $p<0.000001$). However, in this second analysis the test of unidimensionality is passed (95% confidence interval from 4.43 to 11.15). The reliability decreased a little (from 0.911 to 0.900), but this is to be expected since response dependency, which tends to be reduced as items are grouped, artificially inflates reliability (Wainer & Thissen 1996).

These results suggest that when the inter-item dependencies are accounted for (via grouping into testlets based the common EMQ-themes) misfit to the model remained, but there is no longer evidence of multidimensionality.

3. Grouping items into testlets based on specialty

A third and final Rasch analysis was carried out, grouping items by specialty into 18 'super'-items (two testlets per specialty per paper resulting in six in each paper; again see Table 1 for the structure of the exam). This approach produces fit to the Rasch model – see Table 4 (third row) (p=0.289), with none of these 18 'items' showing individual misfit. There are no significant correlations in the residuals and, again, there is no evidence of multidimensionality (95% confidence interval from 2.54 to 8.24).

Summary of Rasch analysis

The proceeding three levels of Rasch analysis have demonstrated that:

- The exam is well-targeted at the students
- Initially, there is misfit to the Rasch model and evidence of both local dependence and multidimensionality.
- However, when the inter-item dependencies are accounted for through grouping items into testlets based on speciality, Rasch model assumptions are met, the data fits the model, and the test for unidimensionality is passed.

TABLE 4 HERE

Discussion

Standard setting using the modified Ebel method

The analysis of the relationship between Ebel judgements and actual performance of items shows that there is a pattern of difference between examiner ratings of question difficulty and actual performance of items/students. There are two components to this difference, the first systematic, where examiners tend to under-estimate the difficulty of items classified as relatively easy under Ebel, and over-estimate that of those classified harder. The second component is a more random one, where at each level of actual item facility there is a spread of difference between the actual item facility and examiner predictions of it (Figure 2).

This second component especially brings into question the comparability of the standard of the examination year on year as set by Ebel-type methods. If Ebel judgements are not sufficiently predictively accurate it is possible that the standard of performance required to pass the exam will not be the same year on year. Since this type of standard setting and examination is widely used, both in the UK and elsewhere, this is an important finding, especially as the stakes are high in this context.

The key question for future research is whether (and how) the system as it stands might be further improved to more closely align Ebel-derived ‘predictions’ of performance with actual performance. One way to improve the alignment might be though feeding back actual item performance to the examiners in post-exam review meetings (Cizek & Bunch, 2007, 80). Future work will investigate the extent to which

additional efforts in this direction have had any impact on the accuracy of the predictions. In addition, it is also hoped that further detailed work on the variation of examiner predictions by specialism might offer some way forward in this regard. It might be the case that examiners are not able to give a realistic judgement of question difficulty in a speciality that is not their own, or even that there might be systematic misjudgement within specialties (i.e. experts not being good judges of how easy or difficult their own questions might be for the borderline candidate).

The Rasch analysis

The Rasch analysis has shown that at the level of specialty (18 testlets) the examination fits the Rasch model in all regards, including being unidimensional (Table 4, row 3). It follows that the summing of the individual testlet scores into one total score for a student is justified, but this is equivalent to summing the individual (i.e. ungrouped) item scores. In other words, this multi-specialty examination is coherent and rational as an assessment with a single outcome measure – student total score. Note that these total scores are ordinal, not interval, in nature but under Rasch, student total score is a sufficient statistic (Panayides et al. 2009) for student ability. That is, the raw total score contains all the necessary information to estimate ability, and it is therefore legitimate to be used as the sole measure, in this context, of clinical knowledge. This is an important (and indeed reassuring) finding.

The Rasch analysis has also demonstrated, for example, that the proportion of items and persons that do not fit the Rasch model are both small (3.1% and 2.0% respectively) and that the reliability is reasonably high (0.887). From a Rasch perspective, it would be reasonable to say state that the examination comes out the analysis well in terms of its overall psychometric characteristics.

However, the lack of fit for the ungrouped and grouped-by-theme analyses (Table 4, rows 1 and 2) cannot be entirely ignored. The item and person estimates for these two analyses are problematic as a result of the misfit and cannot be used with any degree of confidence.

Ebel and Rasch together: towards item banking

In high stakes assessments, all pass/fail (and other grading) decisions must be robust and defensible in the sense that any potential challenges to the outcomes from examinees can be overcome. However, this paper has raised questions over the ability of the modified Ebel method to maintain standards over time. It has also shown that whilst the vast majority of items making up the exam were shown to individually fit the Rasch model, there were problems when considering the exam as a whole at the (ungrouped) item level of detail. Whilst Rasch-based techniques might have potential in terms of increasing the robustness of the standard setting procedure, through the construction of a calibrated item bank (Tennant & Conaghan 2007), the items making up the examination have to be shown to fit the Rasch model assumptions **before** these more advanced aspects of Rasch-based techniques can be fully employed.

To be clear, Rasch analysis cannot produce the ‘standard’ for the examination, but through the use of such an item bank, it can be employed to ensure the maintenance of a given standard across different (though linked) exams. However, such a bank can only be legitimately constructed out of items that fit the Rasch model

requirements both individually and collectively. The analysis has shown that this is not yet possible since items did not collectively fit the Rasch model until they were grouped into testlets (first row of Table 4). Further Rasch-based research will therefore be carried out looking at, for example, differential item functioning by student characteristics including gender and native language (Bond & Fox 2007, 92-95).

It is hoped that this additional work, combined with the ongoing detailed item review informed by Rasch outcomes, will result in a general improvement in the psychometric characteristics of the examination, including further improvement to the item fit statistics. This would then allow the construction of such an item bank to begin. Under such a system, the blueprinting structure that Ebel affords (i.e. ensuring that the exam contains an appropriate mixture of core and peripheral topics, and a good range of difficulty) will still be needed to classify newly constructed items.

Acknowledgments

The authors would like thank Professors Bipin Bhakta and Allan Tennant, and Mr Mike Horton of the Academic Department of Rehabilitation Medicine, University of Leeds, for introducing us to Rasch modelling, and for their constant help and advice in our work in this area. They would also like to thank the helpful comments of earlier reviewers.

References

Andrich D, Sheridan BS, Luo G: RUMM2020: *Rasch Unidimensional Models for Measurement,* Perth Western Australia, RUMM Laboratory; 2002.

Bland, J.M. & Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476), 307-310.

Boursicot, K.A.M, Roberts, T.E., Pell, G. 2006. Standard Setting for Clinical Competence at Graduation from Medical School: A Comparison of Passing Scores Across Five Medical Schools, *Advances in Health Sciences Education,* Volume 11, Number 2, 173-183

Bond, T.G. & Fox, C.M., 2007. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* 2nd ed., Mahwah, N.J: L. Erlbaum.

Case SM, Swanson DB. 1998. *Constructing Written Test Questions for the Basic and Clinical Sciences*. Philadelphia, PA: National Board of Medical Examiners.

Champlain, A.F.D., 2010. A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, 44(1), 109-117.

Cizek, G.J., Bunch, M.B. 2007. *Standard Setting* (2nd edition), Thousand Oaks, California, Sage.

Crocker, L., Algina, J. 1986. *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich College Publishers.

Cusimano, M.D., 1996. Standard setting in medical education. *Academic Medicine: Journal of the Association of American Medical Colleges*, 71(10 Suppl), S112-120.

Downing, S.M., Lieska, N.G. & Raible, M.D., 2003. *Establishing passing standards for classroom achievement tests in medical education: a comparative study of four methods.* Academic Medicine: Journal of the Association of American Medical Colleges, 78(10 Suppl), S85-87.

Ebel, R.L. 1972. *Essentials of educational measurement*. (1st edition)., New Jersey, Prentice Hall.

Goldstein, H., 1979. Consequences of Using the Rasch Model for Educational Assessment. *British Educational Research Journal*, 5(2), 211-220.

Fowell, S.L., Fewtrell, R. & McLaughlin, P.J., 2006. *Estimating the Minimum Number of Judges Required for Test-centred Standard Setting on Written Assessments. Do Discussion and Iteration have an Influence?* Advances in Health Sciences Education, 13(1), 11-24.

Hambleton, R.K., 1991. *Fundamentals of Item Response Theory*, Sage Publications, Inc.

Linacre J.M., 1994. Sample Size and Item Calibration Stability. *Rasch Measurement Transactions,* 7(4), p. 328

Livingston S.A., Zeiky M.J. 1982. *Passing scores: a manual for setting standards of performance on educational and occupational tests*. Princeton, NewJersey: Educational Testing Service.

Maccann, R.G., 2009. Standard setting with dichotomous and constructed response items: some rasch model approaches. *Journal of Applied Measurement*, 10(4), 438-454.

Norcini, J.J., 2003. *Setting standards on educational tests*. Medical Education, 37(5), 464-469

O'Neill, T.R., Marks, C.M. & Reynolds, M., 2005. Re-evaluating the NCLEX-RN passing standard. *Journal of Nursing Measurement*, 13(2), 147-165.

Panayides, P., Robinson, C. & Tymms, P., 2009. The assessment revolution that has passed England by: Rasch measurement. *British Educational Research Journal* (iFirst), 1469-3518.

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.

Skakun, E.N. & Kling, S., 1980. Comparability of Methods for Setting Standards. *Journal of Educational Measurement*, 17(3), 229-235.

SPSS for Windows, Rel. 15.0, 2006. Chicago: SPSS Inc.

Streiner, D.L. and Norman, G.R. (2003). *Health Measurement Scales: A Practical Guide to Their Development and Use*, 3rd edition, Oxford Medical Publications, Oxford.

Tennant, A. & Conaghan, P.G., 2007. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism*, 57(8), 1358-1362.

Wainer, H. & Thissen, D., 1996. How Is Reliability Related to the Quality of Test Scores? What Is the Effect of Local Dependence on Reliability? *Educational Measurement: Issues and Practice*, 15(1), 22-29.

Wainer, H. & Wang, X., 2000. Using a New Statistical Model for Testlets to Score TOEFL. *Journal of Educational Measurement*, 37(3), 203-220.

Wainer H, Bradlow E.T., Wang X. (2007). *Testlet response theory and its applications.* New York, New York, Cambridge University Press.

Figure Captions

Figure 1: Pearson correlation between actual item facility (within decile and overall) and the Ebel-determined passing 'score' for the item (n=330)

Figure 2: Bland-Altman plot of actual item facility and Ebel-determined passing 'score' for the item (10^{th} decile students only) (n=330)

Figure 3: Student-item location distribution (n-student=244, n-item=327)

Tables

Table 1

Paper	Format	Number of questions (items)	Number of questions per set	Number of option sets	Number of marks per question	Specialty breakdown of questions
1	EMQ	120	5	24	1	40 Paediatrics, 40 Psychiatry, 40 Primary care
2	EMQ	120	5	24	1	40 Obstetrics and gynaecology, 80 Medical and surgical.
3	Slide (single best answer)	90	1 or 2 per slide	60	1 or 2	15 Paediatrics, 15 Psychiatry, 15 Primary care, 30 Medical and surgical, 15 Obstetrics and gynaecology.
Total		**330**		**108**	**330**	

Table 2

		Difficulty of item			
		Easy	**Moderate**	**Difficult**	**Total**
Relevance of item	**Essential**	14.2	9.1	0.0	23.3
	Important	12.1	55.8	1.2	69.1
	Additional	0.0	5.2	2.4	7.6
	Total	26.4	70.0	3.6	100.0

Table 3

		Difficulty of item		
		Easy	**Moderate**	**Difficult**
Relevance of item	**Essential**	75%	60%	40%
	Important	55%	45%	20%
	Additional	15%	12%	7%

Table 4

Focus of analysis	Residual item fit		Residual person fit		Item-trait interaction (overall model fit - invariance)			Person separation index (reliability)	Unidimensional t-test percentage (95% confidence interval)
	Mean	SD	Mean	SD	Chi-square	df	p-value		
1. Items ungrouped (n=327)	0.016	1.004	-0.129	1.012	1498.872	981	<0.000001	0.91121	10.66 (6.78,14.53)
2. Items grouped into testlets based on a common clinical theme (n=108).	0.157	0.958	-0.076	0.773	472.197	324	<0.000001	0.89554	7.79 (4.42,11.15)
3. Items grouped into testlets by specialty within each paper (n=18).	-0.149	1.018	-0.181	0.902	59.271	54	0.289349	0.88734	5.39 (2.54, 8.25)

Table captions

Table 1: Structure of the examination

Table 2: The distrubution of averaged Ebel difficulty/relevance ratings of items (percentage, n=330)

Table 3: Examiner consensus on percentage of borderline students expected to know the correct answer

Table 4: Summary of separate Rasch analyses at different levels of focus – item, theme and specialty