International Journal of Testing, 18: 366–391, 2018

ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2018.1486316

Routledge
Taylor & Francis Group

Investigating the Comparability of Examination Difficulty Using Comparative Judgement and Rasch Modelling

Stephen D. Holmes, Michelle Meadows, Ian Stockford and Qingping He

Office of Qualifications and Examinations Regulation, Coventry, UK

The relationship of expected and actual difficulty of items on six mathematics question papers designed for 16-year olds in England was investigated through paired comparison using experts and testing with students. A variant of the Rasch model was applied to the comparison data to establish a scale of expected difficulty. In testing, the papers were taken by 2933 students using an equivalent-groups design, allowing the actual difficulty of the items to be placed on the same measurement scale. It was found that the expected difficulty derived using the comparative judgement approach and the actual difficulty derived from the test data was reasonably strongly correlated. This suggests that comparative judgement may be an effective way to investigate the comparability of difficulty of examinations. The approach could potentially be used as a proxy for pretesting high-stakes tests in situations where pretesting is not feasible due to reasons of security or other risks.

Keywords: examination standards, comparative judgement, Rasch modelling

INTRODUCTION

The General Certificate of Secondary Education (GCSE) examinations in England are public examinations taken by students aged 16 years (or year 11). These examinations are academic qualifications awarded in specific subjects such as English, mathematics and the sciences. They are high-stakes for

Correspondence should be sent to Stephen D. Holmes, Office of Qualifications and Examinations Regulation, Herald Avenue, Coventry, CV5 6UB, UK. Email: stephen.holmes@ofqual.gov.uk

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hijt.

Ian Stockford is now at AQA, Manchester, UK

individuals and schools as results are used for a variety of purposes, including certification of individual students, selection of students for further education or training programs, and accountability for teachers and schools. Unlike the examination systems in other countries, GCSEs are provided by several independent assessment providers known as exam boards in England which are regulated by the Office of Qualifications and Examinations Regulation (Ofqual).

The comparability of standards or difficulty in similar examinations between exam boards has been a concern for the regulator, qualification users and the general public, particularly when existing qualifications are reformed or new qualifications are introduced. Because of the high-stakes nature of these examinations and security considerations, there is no pretesting or test equating, so ensuring the comparability of examination standards between exam boards remains a challenging task (Baird, 2007; Pollitt, Ahmed, & Crisp, 2007).

While it is impossible for exam boards to perfectly control the difficulty of question papers, exams provided by different exam boards that test the same subject areas should be as similar as possible in terms of the levels of demand and difficulty. Slight differences in difficulty between examinations can be accounted for during the awarding (standards setting) process in which boundary (or cut) scores are set for grades using expert judgement of students' work and statistical evidence collected from a range of sources (see Baird, Cresswell, & Newton, 2000; Robinson, 2007). Grade boundaries are set such that they are commensurate with question paper difficulty—more difficult papers have lower boundary scores, while less difficult papers have higher boundary scores. However, consistent and large differences in difficulty cannot be accounted for in this way as the mark distributions may be distorted which could make awarding difficult. Moreover, large differences in difficulty between examinations can undermine confidence in the extent to which there are comparable standards over time and across exam boards. Large differences may also have a negative wash-back effect on teaching and learning. Students being assessed with easier papers may potentially experience a more limited syllabus or less stretching education in the subject or students assessed with harder papers may have a less positive experience of a subject because of their inability to answer many questions on the papers.

England's secondary school qualifications are currently going through a process of reform with both the content being modified and the assessments redesigned (DfE, 2010, 2013). From 2015, the reformed Mathematics GCSE has been taught in schools, with the first examinations having been sat in 2017. As part of the process of introducing these reformed GCSEs,

Ofqual carried out a large investigation into comparability of difficulty between sample assessments and past papers produced by the exam boards (Ofqual, 2015).

Evaluating differences in difficulty prior to the live use of assessments is not straightforward. Actual difficulty, as experienced by students, can only be measured through pretesting. As a proxy for this, expected difficulty can be estimated, based on a judgement of the difficulty students might experience when answering items. Techniques such as the Angoff (1971) and bookmark (Cizek & Bunch, 2007) methods, where experts set the cut points for proficiency or other performance standards such as grades, by considering individual item difficulty, show that expert judgement can be used in evaluating item difficulty. These methods only give information on items around the cut points or how a minimally qualified individual would perform on items, not an overall estimate of test difficulty. In order to understand overall test difficulty across the ability range of candidates for whom the test is intended, the expected difficulty of each item is required. Asking experts to independently judge the difficulty of items generates rank order correlation to actual difficulty only of the order of 0.3 (e.g. Bejar, 1983; Melican, Mills, & Plake, 1989). Hambleton, Sireci, Swaminathan, Xing, and Rizavi (2003) found the same correlation for independent judgements and that inclusion of anchor items of known difficulty did not greatly increase the correlation above 0.3. However group discussion slightly improved the rank order correlation to almost 0.5 and providing items from an entire test with known actual difficulty as a framework generated correlations of 0.61 for independent predictions and 0.76 with group discussions. This last approach has more in common with comparative judgement approaches than independent ratings. This suggests that experts do not have a robust internal scale against which to rate items (Hambleton et al., 2003). Therefore a comparative approach is likely to be more successful. Attali, Saldivia, Jackson, Schuppan, and Wanamaker (2014) used a comparative approach with test developers and item writers rank ordering the expected difficulty of random packs of 7 items, and found a mean correlation of 0.7 between rank orders of estimated difficulty and actual difficulty within packs.

Rank ordering becomes increasingly difficult the larger the set of items in question. Bramley (2005) showed that overlapping packs of rank ordered items can be combined into a single rank order, but an alternative way to obtain comparative data is to use the paired comparative judgement approach, which is detailed in Thurstone (1927), Bramley (2007) and Pollitt (2012). To explore this idea, this paper focuses on two major studies included in the Ofqual (2015) investigation which allowed comparison of

expected and actual difficulty. The first study explored the comparative judgement approach to predicting expected examination difficulty. This was used to compare the expected difficulty of mathematics items (questions) drawn from sample assessments and recent past papers from the exam boards and a selection of equivalent papers from international jurisdictions. In the second study, six of the mathematics question papers from the exam boards were taken by 2933 16-year old students using an equivalent-groups design with the analysis of the test data based on the Partial Credit Rasch Model (PCM) (Masters, 1982; Wright & Masters, 1982) to obtain the actual difficulty of items and papers. The relationship between the expected item difficulties estimated using the comparative judgement approach and actual difficulties obtained from the test data using the PCM model was examined in order to assess the validity of using comparative judgement as a new approach to predicting examination difficulty. The present article reports the main findings from the two studies.

RASCH MODELLING AND COMPARATIVE JUDGEMENT

The Rasch Models

The Rasch family of models have been used extensively to analyze data from psychological and educational tests to establish measurement scales. In Rasch modelling, the underlying ability or latent trait of an examinee to be measured by the test and the characteristic of the items to be measured in the test (usually difficulty) are specified, and a mathematical function is used to describe the probability that a person will have a specific score on a particular item given his/her ability and the difficulty of the item. The Rasch model for dichotomous items can be expressed mathematically as (Rasch 1960; Wright & Stone, 1979):

$$P(\theta) = \frac{\exp(\theta - b)}{1 + \exp(\theta - b)} \tag{1}$$

Where θ is the ability of the person, b is the difficulty of the item, and $P(\theta)$ is the probability that the person will answer the item correctly. It can be seen from Equation (1), when the item difficulty is close to person ability, the test-taker will have a 50% chance of answering the item correctly. The dichotomous model has been extended subsequently for analyzing polytomous items (items with two or more marks available). These extended Rasch models include Andrich's Rating Scale Model (RSM), Masters' partial credit model (PCM), and other models (see Andrich, 1978a; Masters, 1982; Wright & Masters, 1982; Muraki, 1992). In the PCM, the probability $P(\theta, x)$ of an

examinee with ability θ scoring x on a polytomous item with a maximum available score of m can be expressed as:

$$P(\theta, x) = \frac{\exp \sum_{k=0}^{x} (\theta - \delta_k)}{\sum_{l=0}^{m} \exp \left[\sum_{k=0}^{l} (\theta - \delta_k) \right]}, \quad (2)$$

where δ_k is the location of the k^{th} score category on the latent trait continuum and is referred to as the item category parameter associated with a score category (also frequently referred to as category threshold, see Andrich, 2015). However, δ_k cannot be interpreted as the difficulty of obtaining a score of k on the item (see Wu & Adams, 2007). $P(\theta, x)$ is also frequently referred to as category response function (CRF). The category parameter δ_k represents the location of the score category on the ability continuum beyond which the probability of achieving a score of k is higher than that achieving a score of k - 1. The partial credit model was used to analyze the test data from the six papers.

Comparative Judgement

Louis Thurstone proposed the paired comparison method for placing psychological phenomena on a chosen psychological continuum (Thurstone, 1927). Pairs of stimuli are presented and the one rated to have more of the quality of interest is chosen. It is based on the observation that people are better at making relative judgements than absolute judgements (e.g., Laming, 1990). The application of the paired comparison approach in educational studies is described in Bramley (2007) and Pollitt (2012).

Andrich (1978b) demonstrated that Thurstone's special case V of the paired comparison model can be closely approximated by the Rasch logistic model. This is also known as the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce 1959), which we used when fitting the comparative judgement data. If a judge is asked to make a paired comparison of a property of two objects A and B with the property values of υ_A and υ_B respectively, the probability that A beats B can be expressed as:

$$P(A>B) = \frac{\exp(\upsilon_A - \upsilon_B)}{1 + \exp(\upsilon_A - \upsilon_B)} \quad (3)$$

As can be seen Equation (3) takes the same form as the Rasch model for dichotomous items represented by Equation (1). The Bradley-Terry-Luce model was used to establish the scale for the predicted difficulty of the items

in the six papers studied. Note that this difficulty scale is a relative scale, with position along the logistic scale indicating the log odds of one item being judged more difficult than another item. The absolute value is arbitrary, in this case 0 is set equal to the mean of all the items included in the study. The values therefore carry no information about the absolute difficulty of these items for candidates.

METHOD

Items and Question Papers Studied

GCSE mathematics question papers are tiered for difficulty, with the most able students taking higher tier papers and less able students taking foundation tier. We report on the difficulty of items on three foundation and three higher tier papers produced by three exam boards for the purpose of accreditation of the reformed qualifications. These six papers are a subset of those used in the larger comparative judgement and test studies (Ofqual, 2015), but are the only papers included in both studies. The exam boards are labelled A, B and C respectively in this study. The number of items in each of the papers and the maximum raw and scaled marks are listed in Table 1. The use of scaled mark is to make the comparison between the question papers more meaningful.

Comparative Judgement of Expected Item Difficulty

Questions from the papers were divided into the smallest question parts to form individual items for the comparative judgement study. Mark schemes were not included, to encourage the judges to work through the items and uncover any unexpected difficulty. Additional items from past GCSE maths papers and assessments from several international jurisdictions were also used in order to investigate the comparability of difficulty between the existing

TABLE 1
Number of Items, Marks in the Six Question Papers, and Number of Students Taking the Papers

Paper	Number of Items	Maximum Raw Mark	Scaled Maximum Mark	Number of Students
Foundation Paper, Board A	44	80	100	325
Foundation Paper, Board B	48	100	100	362
Foundation Paper, Board C	37	80	100	353
Higher Paper, Board A	37	80	100	618
Higher Paper, Board B	36	100	100	648
Higher Paper, Board C	28	80	100	627

GCSEs, the reformed GCSEs and international examinations. However, this is not reported in this article.

The model represented by Equation (3) is implemented in the online system *No More Marking*[1] for paired comparison which was used for comparing the difficulty of items studied here. Online systems allow judgements to be made in a distributed fashion with large numbers of judges, which has the further advantage of cancelling out individual bias. Once enough judgements have been made a scale can be created from the judgements using the Bradley-Terry-Luce model. The construction of a scale allows properties of the model, such as the consistency of judgement and the reliability of judgement, to be evaluated. There is growing evidence that comparative judgement can establish robust scales that are validated by external criteria (e.g., Jones, Swan, & Pollitt, 2015).

The items were loaded into the online system and pairs of items were judged by a group of 43 PhD maths students from English universities for expected relative difficulty. Judges were given detailed instructions on how to access the platform and how to make their judgements. Pairs of items were presented side by side on the screen and the judges were prompted to indicate: "Which Item is the More Mathematically Difficult to Answer Fully?"

The judging prompt was always present on the judging screen. The judges were specifically asked to judge the mathematical difficulty of the items. Each item was judged (compared with other items) about 30 times. The expected relative difficulties of the items were then estimated using the Bradley-Terry-Luce model discussed previously.

Investigation of Actual Item Difficulties

In the second study, the difficulty of the items in the question papers was investigated. Year 11 students preparing for their Maths GCSE from a selection of schools were recruited to take the question papers as a mock examination. The schools were chosen to be representative of the national sample based on their size, prior attainment of their current student cohort, and the percentage of students in their most recent cohort achieving five or more passing GCSE grades (including English and mathematics). However there was no need for the sample to be perfectly representative of the national sample as the relative performance across papers was more important than the absolute performance. Each school selected the entry tier (higher or

[1] www.nomoremarking.com.

foundation) for their students and provided this information on their candidate lists. The exam boards' question papers were randomized within each class within each school to ensure that the groups taking each paper in a tier were randomly equivalent. This equivalent-groups equating design makes it possible to place items in the papers on the same scale. The sample size required was estimated by the precision of relative item facility that could be achieved—a target sample of 500 participants per paper would give a 95% confidence interval of ±0.04 of true item facility. The final achieved number of students sitting each of the papers is listed in Table 1. In total, 2,933 students took the six papers.

Responses from the students on the items were marked online by trained experienced markers using standard procedures. The marks were then analyzed to obtain item difficulty parameters that were aggregated to whole assessment level using the PCM implemented in the Winsteps Rasch analysis software (Linacre, 2017). The actual difficulty of the items derived using the PCM were then compared with the expected difficulties predicted using the comparative judgement approach.

RESULTS

Expected Item Difficulty Predicted from the Comparative Judgement Method

Fit of the Comparative Judgement Model. Eight judges were excluded from the analysis on the basis of the haste and lack of consistency with which they made judgements. The eight judges had a median judgement time of less than 10 seconds per paired comparison and outfit values that ranged from 1.08 to 1.48 (for an explanation of outfit, see, Pollitt, 2012). The range of outfit values for the remaining 35 judges included in the analysis was 0.81 to 1.10, while the range of median judgement times by judge was 11 to 35 seconds with a mean of 19 seconds.

Reliability is quantified in CJ studies by the scale separation reliability (SSR) statistic that is derived in exactly the same way as the person separation reliability index in Rasch analyses. It is interpreted as the proportion of "true" variance in the estimated scale values. Once the eight judges were removed, the SSR was 0.88. The reliability values suggest a certain degree of disagreement among the judges, but not enough to threaten the measurement properties of the expected difficulty scale created.

Expected Item Difficulties. Figure 1 shows the median and distribution of the item expected difficulties for each of the six papers based on comparative judgement. No weighting by tariff (maximum mark on the item) has been

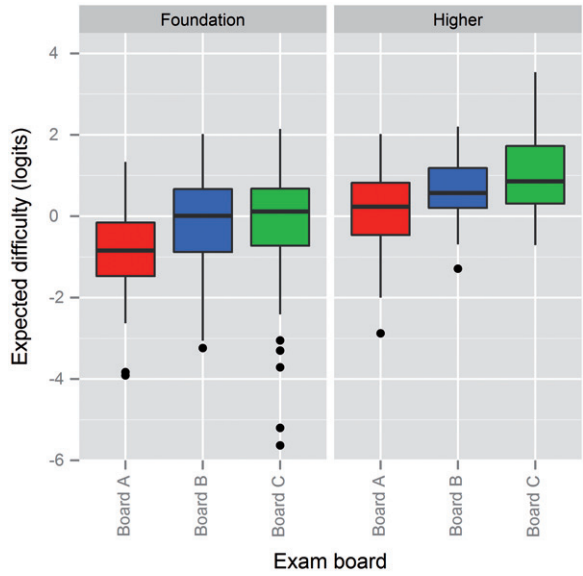

FIGURE 1
The median and interquartile ranges of expected item difficulty parameters for the six question papers studied.

applied to this data. Differences across the papers from the exam boards are apparent, but as the expected difficulty scale is arbitrary the operational significance of these differences is not clear in terms of actual mark distributions. However it is noteworthy that the Board A higher tier paper is almost indistinguishable in difficulty from the Board B and C foundation tier papers, which are intended to be targeted at a different ability range. It is also worth noting the degree of overlap between the papers from each exam board on the different tiers. Board B's papers appear to have greater overlap than the other two boards. Although some overlap would be expected, as there are some common grades across the tiers, there is not much difference in the upper range of the foundation and higher-tier papers.

Relationship Between Item Tariff and Expected Difficulty. Figure 2 shows that there is a slight relationship between expected item difficulty and tariff. For the foundation tier, 1-mark questions include some that are judged less difficult, while most of the other questions do not show much of a trend. For the higher tier, there is more of a trend; items with high tariffs are generally judged to be more demanding than items with low tariffs. This effect of tariff may have been enhanced by the judges evaluating the difficulty

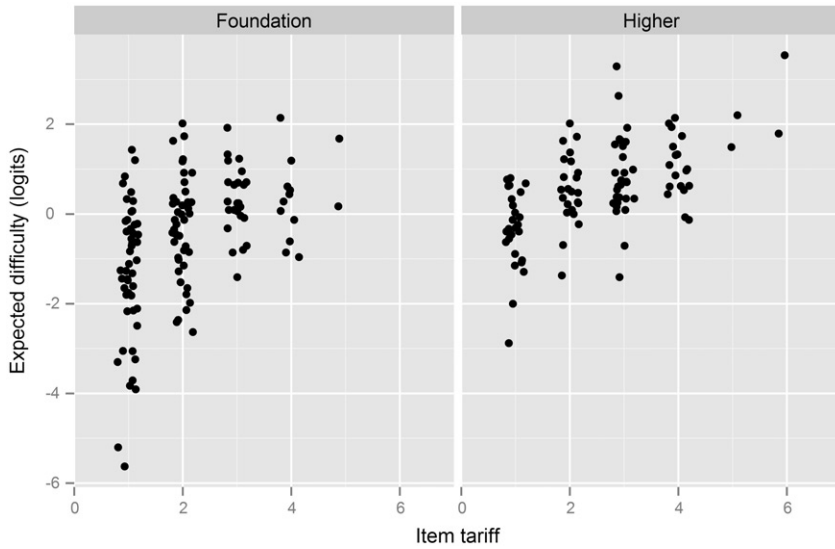

FIGURE 2
Comparative judgement expected item difficulty by item tariff (mark).

of giving a full (i.e., full-mark) answer. However, this relationship is not particularly strong.

Expected Test Characteristic Curves. To make the comparison of difficulty between the question papers more intuitive, the predicted relative difficulties of the items were used to derive a test characteristic curve (TCC) for each of the six papers using the approach proposed by He and Wheadon (2013), which uses the Rasch model for dichotomous items to analyze polytomous items. In this approach, for a polytomous item i with a maximum score of m_i, the overall difficulty of the item is approximated by one difficulty measure δ_i, and the attempt by person n with ability θ_n at the item is assumed to be similar to answering m_i dichotomous items with equal difficulty δ_i. Under this assumption, the dichotomous Rasch model used for analyzing dichotomous items can be extended for analyzing polytomous items. When the average difficulties of the items in the test are known, the TCC can be defined as $\Gamma(\theta) = \sum_{i=1}^{I} m_i P(\theta, \delta_i)$, where I is the total number of items in the test. By treating the predicted relative item difficulties as the actual item difficulties, the TCCs for the six papers were derived which are shown in Figure 3. The test characteristic curve shows the relationship between the expected score on the test and person ability. When the curves for different tests are placed on

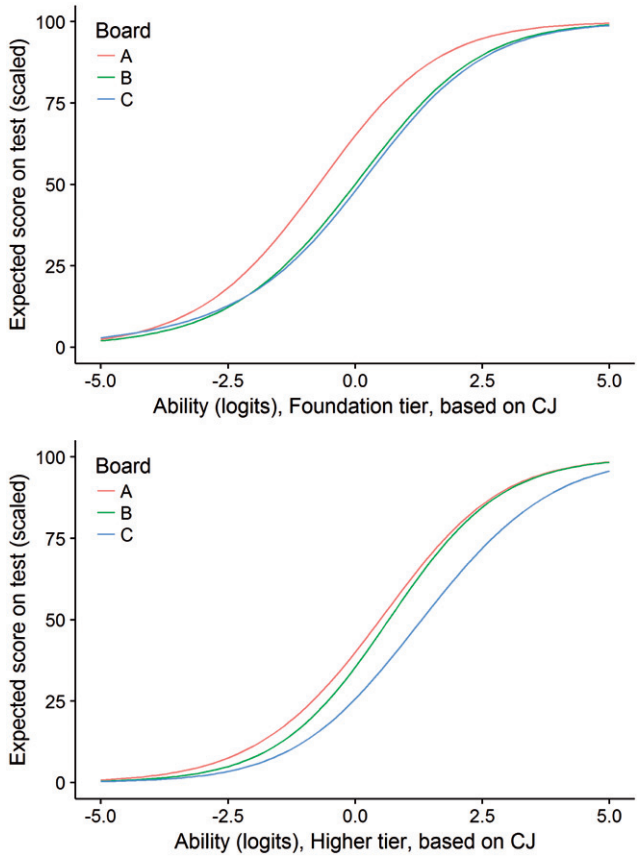

FIGURE 3
Test characteristic curves for the six papers derived using the expected item difficulties estimated using comparative judgement and the Rasch model for dichotomous items.

the same ability scale and have the same shape, the test on the left will be easier than those on the right, since for the same ability the expected score on the test will be higher than those on the other tests.

These curves reflect the differences shown in Figure 1, including capturing the relatively reduced content at the top of the ability range for the Board B higher tier paper, which converges with the difficulty of the Board A higher tier paper for the most able students. For the foundation tier, the papers from Boards B and C performed similarly and were of similar difficulty across the full ability range. The paper from Board A was however considerably easier than the other two papers. For the higher tier, the expected scores from the

paper from Board C were considerably lower than those from the other two papers across the ability range. It is unclear whether there is sufficient separation of the papers between the two tiers for Board B. Using the 0-logit ability line as the reference, for Boards A and C there is a difference of approximately 25 scaled marks between papers on the two tiers. For Board B this difference is only 15 scaled marks.

These TCCs suggest that the maximum difference between papers from different exam boards within a tier is around 15 scaled marks. This is quite a substantial difference in difficulty between the papers. If the ability of the student cohort was matched to the difficulty of the papers (e.g., centered around 0.0 logit for foundation tier or 1.0 logit for higher tier) the mark distributions for all the papers may be acceptable, but any mismatch between cohort ability and paper difficulty could lead to one or more papers showing an excessively skewed mark distribution, which would make awarding difficult.

Item Difficulty Derived from the Test Data

Classical Test Theory Analysis. All candidates sitting the papers were included in the analysis. The number of students with a zero score on the Foundation papers was 11 for Board A and 12 for Boards B and C, while that for the Higher Tier papers was 7 for Boards A and B and 23 for Board C. These are small. The performance of students on the six papers is summarized in Table 2. For the purpose of comparison, all papers were scaled to have a maximum available score of 100. All papers had good internal consistency reliability, with the lowest value of Cronbach's alpha at 0.83. The scaled mean scores of students on the papers from Board A were higher than those of students sitting the other boards' papers. This is illustrated graphically in Figure 4, which shows the distribution of student marks on the papers. The differences in difficulty between exam boards were statistically significant.

For all the papers, the items were generally too difficult relative to the abilities of the students from a measurement perspective, which may partly reflect the effect of motivation associated with pretesting as these papers were taken under low-stakes conditions (see Pyle et al., 2009; Steedle and Grochowalski,

TABLE 2
Question Paper Analysis

	Board A		Board B		Board C	
	F	H	F	H	F	H
Mean scaled score	40.16	33.91	24.44	27.98	23.62	18.04
Scaled standard deviation	18.53	16.89	15.82	16.96	13.14	13.01
Cronbach's alpha	0.90	0.88	0.87	0.88	0.84	0.83

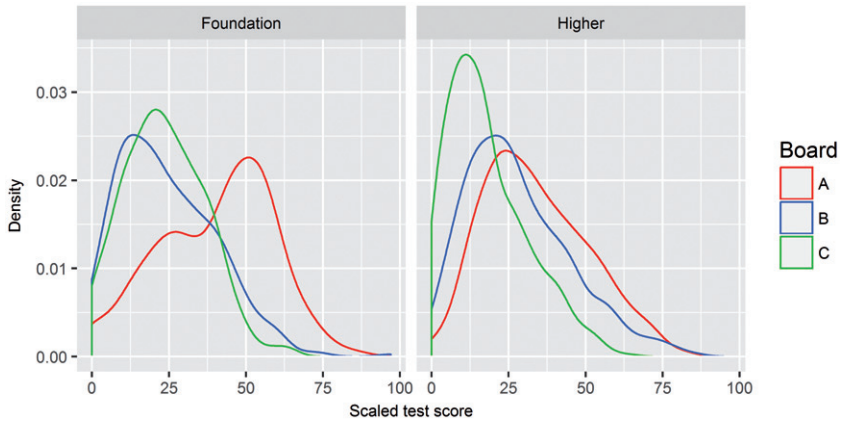

FIGURE 4
Scaled score distributions for the six papers.

2017). Test-taking motivation can affect test performance considerably. In a low-stakes testing environment, where there are little or no consequences associated with test performance or perceived benefits for the test-takers, the performance on a test can be considerably lower than the performance under high-stakes conditions. For example, in the study on the effect of motivation on the performance of the Key Stage 2 National Curriculum Science tests that were being pretested under a low-stakes condition, the overall pretest effect represented an increase in average test facility of 14% in 2006 and 13% in 2007 when comparing live testing with pretesting (Pyle et al., 2009).

Unidimensionality and Model-Data Fit. The test data from the six papers were analyzed using the partial credit Rasch model. The appropriateness of the use of the PCM is dependent on the data being unidimensional and fitting the model. Unidimensionality requires that one ability or a single latent variable is measured by the items in the test. The unidimensionalilty assumption of the model also requires that test-takers' responses to any questions in the test are statistically independent when their underlying ability influencing their performance on the whole test is held constant (local independence). The extent to which items in a test meet model assumptions needs to be investigated, as their violation can invalidate the interpretation of results. Unidimensionality and local independence can be evaluated using factor analysis of row scores or the residuals of person scores (defined as the differences between observed scores on items and Rasch model predicted scores) and the analysis of residual correlation matrices between items (see Lord, 1980; Hambleton, Swaminathan and

Rogers, 1991; Yen, 1993; Yen, Bené, & Huynh, 2000; Smith, 2002; McGill, 2009; Reckase, 2009; Linacre, 2017). The application of the Rasch model also requires that test data fits the model sufficiently well. The degree to which the data fits the model can be evaluated using a range of model fit statistics such as residual-based unweighted mean squares (outfit MNSQ) and weighted mean squares (infit) (see Wright & Masters, 1982; Wu & Adams, 2007; Linacre, 2017).

Table 3 shows the ratio of the first eigenvalue to the second eigenvalue from exploratory factor analysis (EFA), the variances explained by the Rasch model, standardized root mean square residual (SRMR), correlations of residuals between items, item fit statistics, and Rasch person separation index and reliability for the six papers. The ratio of the first to the second eigenvalues is substantially higher than the ratio of the second to the third eigenvalues for the papers, suggesting that there is a strong common dimension running through all the items in the papers. The percentage of variance explained by the Rasch model varied from 49% to 63%, indicating that the tests broadly loaded on a single dimension. Principle components analysis (PCA) of residuals indicated that the ratio of the first contrast to the second in the residuals in eigenvalue terms varied from 1.03 for the Foundation Tier paper from Board A to 1.28 for the Higher Tier paper from Board C which is similar or slightly higher than the ratio of the second contrast to the third contrast. This suggests that these contrasts are of relatively equal importance in explaining the variance not explained by the Rasch model and therefore it may be assumed that no meaningful second dimension could be constructed for the original test data. The correlations of residuals between pairs of items varied from -0.23 for the Foundation paper of Board B to 0.34 for the Higher paper from Board C, with most of the item pairs in the papers having absolute values of correlation of residuals less than 0.20, suggesting that the dependence between items is insignificant. The values of the SRMR for the tests varied from 0.05 to 0.2, which suggests good overall model-data fit (see Maydeu-Olivares, 2013). At individual item level, for most of the items, the value of infit is above 0.78 and below 1.35, with only two of the items from the two Higher Tier papers from Boards A and C having infit slightly above 1.50. Except for the Higher tier paper from Board C, all the other papers had Rasch based person reliability above 0.84. It is noted that views about the criteria that should be used to evaluate model assumptions and model fit vary among researchers and such criteria generally depend to a large extent on the purpose of the investigation (Hambleton, et al., 1991; Linacre, 2002; Tan & Yates, 2007; McGill, 2009; Pae, 2012; He, Anwyll, Glanville, & Opposs, 2014; Christensen, Makransky, & Horton, 2017). Further, as suggested by Hambleton, et al., (1991), as long as a coherent scale can be constructed by the items, strict unidimensionality will not be needed because IRT analysis is

TABLE 3
Dimensionality, Model Assumptions and Model Fit

	Board A		Board B		Board C	
	F	H	F	H	F	H
Ratio of first eigenvalue to second eigenvalue from EFA	3.7	4.5	4.2	4.3	3.6	2.9
Variance explained by Rasch model (%)	54.3	60.2	49.8	58.5	60.8	48.8
Ratio of first contrast to the second contrast	1.04	1.14	1.17	1.26	1.04	1.27
Standardized Root Mean Square Residual (SRMR)	0.05	0.02	0.04	0.02	0.04	0.02
Correlations of residuals between items						
Mean	−0.02	−0.02	−0.03	−0.03	−0.02	−0.04
Range	−0.22-0.27	−0.18-0.22	−0.23-0.22	−0.21-0.19	−0.17-0.34	−0.21-0.10
Standard deviation	0.07	0.06	0.08	0.06	0.07	0.06
Mean of infit	1.00	1.00	1.00	1.02	1.00	0.99
Range of infit	0.82-1.24	0.78-1.51	0.81-1.18	0.81-1.35	0.83-1.35	0.80-1.56
Person separation index	2.92	2.60	2.42	2.49	2.30	1.83
Person reliability	0.90	0.87	0.85	0.86	0.84	0.77

relatively robust to violations of the unidimensionality assumption (also see Bèguin 2000; Hanson & Bèguin, 2002). Given the exploratory nature of this study, these results indicated that the assumptions of Rasch model were met and that the data fitted the model sufficiently well for the purpose of the present study.

Observed Item Difficulties. Figure 5 compares the distribution of the difficulties of the items (calculated as the mean of category thresholds for individual items) and the distribution of person ability for the six papers (the Wright Maps). For each of the two tiers, the distribution of person abilities for the three papers were similar, which suggested that the assumption of equivalency in ability between the groups held reasonably well. Compared with

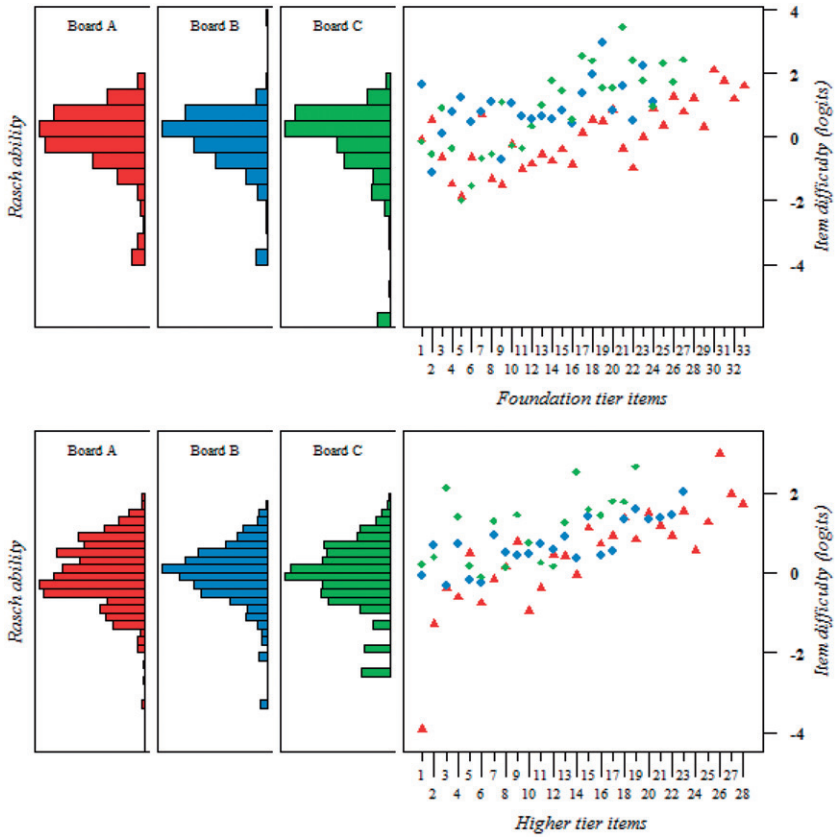

FIGURE 5
Distributions of person abilities for the 2933 students taking the six tests and the average item difficulties of the six tests (Wright Maps).

Figures 4 (the scaled score distributions), Figure 5 indicates that the distribution of pupils on the Rasch ability scale is more symmetric than on the scaled score scale. This is because although scores are sufficient statistics for estimating the item parameters of the Rasch model, the Rasch ability is not a linear function of scores—pupils with extreme scores were stretched outwards. The use of the Rasch scale therefore removes the floor and ceiling effects associated with scores. The item difficulty generally increases with item order. This is particularly so for Board A, which has very clear ramping of difficulty through the papers, except for a few foundation tier items early in the paper. Information contained in graphs like Figure 5 can be used to improve test construction. In the case of large gaps between item difficulties in the ability scale, students with abilities close to the gaps will be less precisely measured. The difference in difficulty between papers seen in Figure 4 is reflected in Figure 5, and for all the papers, the questions were generally too difficult relative to the abilities of the students from a measurement perspective.

Figure 6 shows the median and distribution of the actual item difficulties for each of the six papers based on the Rasch analysis. No weighting by tariff has been applied to this data, and the difficulty scales of the two tiers are not

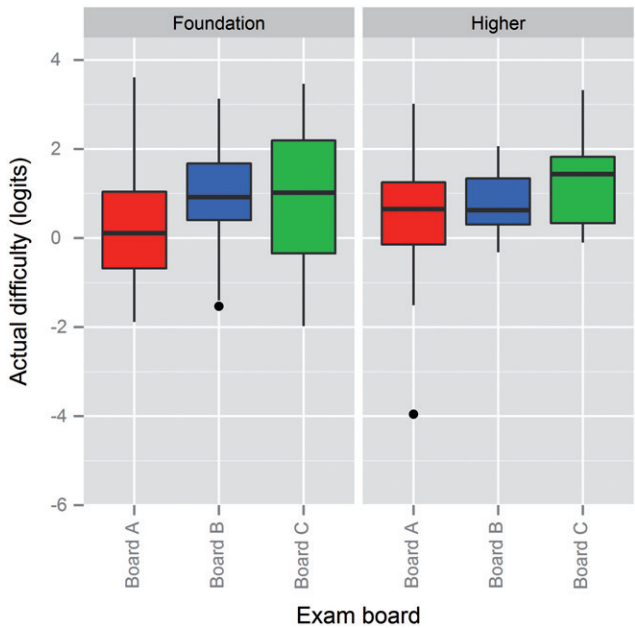

FIGURE 6
The median and interquartile ranges of actual item difficulty derived from the PCM and test data for the six question papers studied.

equated, but reflect the independent application of the PCM model within each tier, with the cohort ability centered on 0 logits. There is reasonable consistency with the expected difficulty distributions seen in Figure 1. The relative difficulties of the foundation tier papers are reproduced across the two methods quite well. For the higher tier papers the main difference lies in the Board B paper, which does not appear any more difficult than the Board A paper seen here. Both Figures 1 and 6 capture the narrow range of difficulty of the Board B higher tier paper.

Relationship Between Item Tariff and Actual Difficulty. Figure 7 shows the relationship of actual item difficulty to mark tariff, split by tier. Unlike the data shown in Figure 2, there does not appear to be any relationship between difficulty and item tariff. For students, the difficulty of a question is not related to its tariff, while for our judges in the comparative judgement study there was a weak relationship. This difference may come from the judgement made by the judges, where the difficulty of giving a full answer (which would imply one that is awarded all available marks) was judged. Although the item tariffs were not presented in the comparative judgement study, tariff relates to the complication or number of steps in the question, and so judges would have rated high-tariff questions harder to give a full answer on than low-tariff questions.

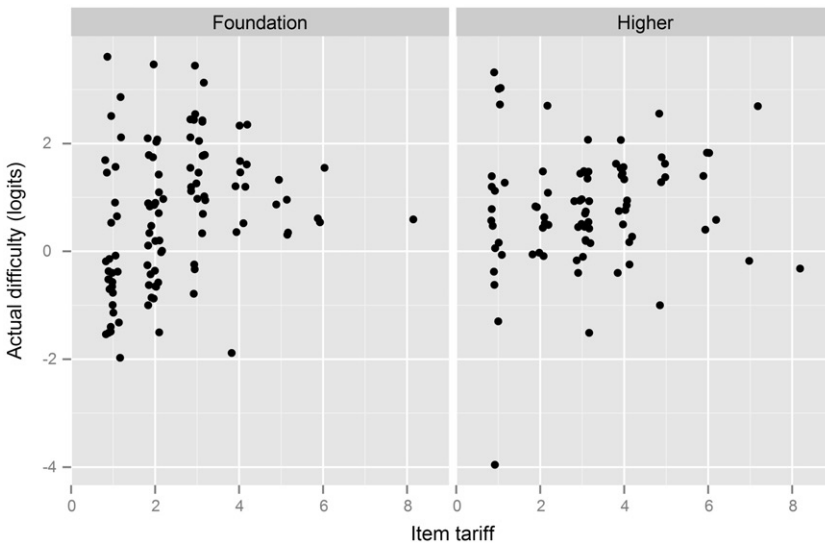

FIGURE 7
Actual item difficulty derived from the PCM and test data by item tariff (mark).

Test Characteristic Curves. Figure 8 compares the TCCs of the three papers from each of the two tiers. Again, the available marks on the papers were scaled to have a maximum of 100. As the average ability for each of the groups was set to 0, the expected scores for persons with an average ability on the tests vary and are well below 50 marks for all six papers (half of the maximum available scaled marks on the papers). For the foundation tier papers, the paper from Board A is considerably easier than the other papers across the full range of ability. For the other two papers, their difficulties vary with ability. When ability is below about 0.2 logits, the Board B paper is slightly harder than the other two papers. From the ability of 0.5 logits to 2.5 logits, the paper

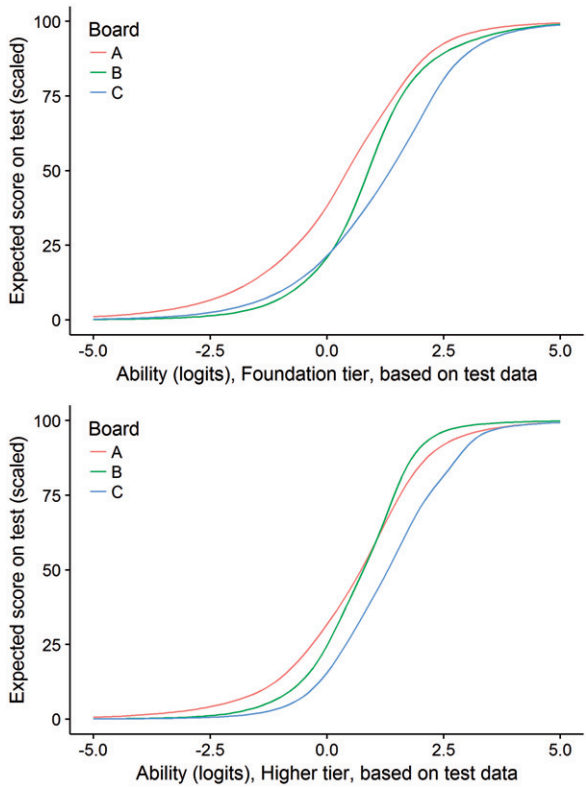

FIGURE 8
Test characteristic curves derived based on students' responses to items and the Partial Credit Model for the six papers studied (For each tier, a random equivalent-groups design was used. Therefore the person ability distribution for the three groups was the same and the items in the papers were on the same difficulty scale).

from Board C is more difficult than the other two papers. For the Higher Tier papers, the paper from Board C was more difficult than the other papers. The Board A paper was slightly easier than the Board B paper below the ability of about 1.0 logits but slightly harder than the Board B paper above 1.0 logits.

The similarities and differences between the exam boards found in the comparative judgement data (Figures 1 and 3) are evident in comparing these functions to Figures 6 and 8. The relative difficulties of the papers are reproduced to a first approximation between the two methods, including the range of difference in scores between papers within a tier (e.g., up to 15–20 scaled marks). However, the performance of the Board B papers does differ across methods. The test TCCs show a more pronounced difference in the spread of abilities catered for, with the Board B papers at both tiers being more narrowly targeted along the ability scale. While the comparative judgement TCCs suggested that the Board B foundation tier paper should function almost indistinguishably from the Board C foundation tier paper, the test data showed that for the most able students it approached the difficulty of the Board A paper. The application of an analogue of the dichotomous Rasch model to the polytomous items for the comparative judgement data could also have contributed to the differences in TCCs between the two approaches. Slightly different curve slopes and different relative difficulty of papers for different student ability could also have been caused by the way intermediate marks are awarded, which are not picked up by the judgement of full-mark difficulty collected in the comparative judgement study.

Comparison of CJ Predicted Item Difficulties and Observed Item Difficulties

Figure 9 shows the relationship between the item difficulty parameters derived using comparative judgement and the actual difficulties derived from the test data using the PCM. To produce this figure, the actual item difficulties for the foundation and higher tier were placed onto a single difficulty scale through the equating of the common items across the two tiers. The relationship between expected and actual difficulty is reasonably strong, with a correlation of 0.67 and a disattenuated correlation (taking account of measurement error in the estimates of item difficulty) of 0.76.

DISCUSSION

For the question papers investigated, there was good consistency between the expert judgements and actual test administration approaches to estimating item difficulty. Certainly, the correlation obtained here was of a similar magnitude

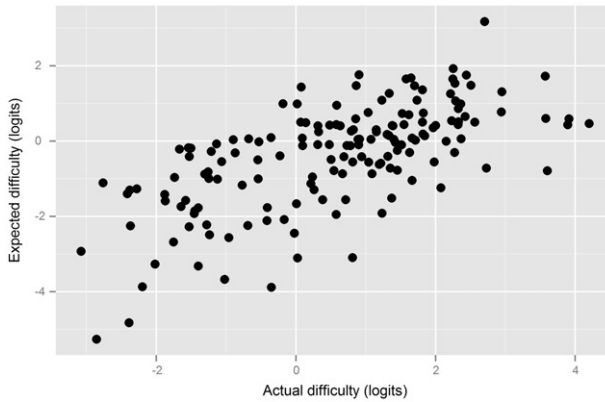

FIGURE 9
The relationship between expected item difficulty derived using comparative judgement and the actual item difficulty estimated based on students' responses using the PCM for the items contained in the six papers.

to previous efforts to judge difficulty using anchor items or anchor tests (e.g., Hambleton et al., 2003). Although difficulty estimates of individual items show some spread, when aggregated across whole papers the predicted and actual TCCs were broadly similar. When thus combined, providing there are no systematic biases in judging expected difficulty of items from different exam boards, the median and spread of predicted item difficulty for a paper will represent the actual difficulty of that paper reasonably well.

Given this, the comparative judgement approach could be used in place of pre-testing to compare the difficulty of future live papers. Where large differences are found, expert judgement in replacing items could be combined with an iterative comparative judgement process to adjust test difficulty. In practice, in the English national testing context this would be logistically difficult. It is also worth noting that the regulator and the exam boards also use well established methods (e.g., Ofqual, 2017) to ensure the comparability of standards between exam boards when grading live question papers, which also address difference in difficulty of assessments between exam boards. However comparison of papers across years could be done using this method to increase consistency of difficulty. This would be consistent with the need to always include past-papers/assessments of known difficulty with which to benchmark the new assessments because the comparative judgement approach does not give an absolute estimate of difficulty—in this study it was unable to reveal that several of the papers were likely to be too difficult for the target cohort.

The differences between expected and actual difficulty, represented by a correlation somewhat less than 1, may come from several sources. Underlying any difference may be differences in the nature of the conceptualization and operationalization of difficulty between the two approaches. The actual difficulty of the papers may vary according to how they are operationalized. For example, a seemingly challenging but predictable assessment can be very easy for students to complete. There is also the known problem of experts with advanced subject knowledge trying to evaluate difficulty for students with much less knowledge (Pollitt et al., 2007). For example, Shermis and Chang (1997) showed that for mathematics items even in cases where empirical difficulty data and expert judgement are in good agreement, there are anomalies, and experts may misjudge the difficulty for candidates of certain types or categories of items. This problem cannot be eliminated by the comparative judgement approach, particularly if all of the judges share the same biases. Although PhD students are less remote in time from the act of taking these exams than assessment experts may be, they are highly advanced mathematicians. It was for this reason that the judges were asked to rate the underlying mathematical difficulty of the questions, rather than the difficulty a student would have in answering. However this could have led to some misalignment between their estimates and the actual difficulties. In the papers under consideration here this misalignment is unlikely to have much of a differential effect across papers. There are tight constraints upon the proportion of marks across different maths domains and assessment objectives (learning outcomes) that each board must satisfy, and so any systematic biases would be almost equally distributed across boards.

Further difference comes from the PCM fitted to the test data giving an overall measure of difficulty, taking the difficulty of obtaining each of the marks into consideration, compared to the Bradley-Terry-Luce model fitted to the comparative judgement data, which gives a measure of the difficulty of achieving full marks, not overall difficulty. This difference may have led to the slightly different slopes of the TCCs for the Board B papers (particularly the Foundation tier) across the two methods in the present study. Any differences across papers in the difficulty of achieving intermediate marks for items could lead to slightly different slopes. For example, a number of items on a test for which the intermediate marks were very difficult to obtain and where the item characteristic curve was therefore very steep would narrow the TCC for actual test administration relative to the CJ approach, where this difference would not be represented. Although this small effect appears to have been present here, in general, when predicted item difficulties are aggregated together at paper or assessment level, the reasonably strong relationship between the comparative judgement expected difficulties and the observed actual item

difficulties suggests that comparative judgement can be effective for predicting examination difficulty.

Given such limitations, there is scope for improving the predictive power of the comparative judgement approach. First, judges could use their expertise to consider the difficulty a student of the appropriate age would have in answering the question, to capture some of the non-mathematical demands that were excluded in this study, while acknowledging the potential limitations discussed earlier in the accuracy of this judgement. Second, the criteria against which judges compare the pairs of items could also be brought into closer alignment with the measure of difficulty captured by the PCM difficulty estimate. In a paired comparison, judging which item a student would be expected to gain a greater proportion of the maximum marks could be made. However, this kind of judgement may be harder for judges to make than assessing the difficulty of giving a complete correct answer, and it may lead to within- and between-judge inconsistency where a judge's criteria shifts over time, or each judge has a slightly different understanding of the criteria. Judging the difficulty of giving a completely correct answer, as used here, had the benefit of a very clearly defined benchmark against which to judge. An additional shortcoming of judging against some notion of overall difficulty is how to fit a model to the data to start estimating TCCs and the like, given that the initial model fit gives only a single parameter that is located at an ill-defined point on the item characteristic curve (ICC). Further work on the basis upon which judges made their judgements in comparative judgement studies could also provide useful information to improve the comparative judgement approach.

The use of expert judgement cannot address all aspects of pre-testing for all tests. In general test development and construction it does not offer the ability to detect differential item functioning or to remove and replace other poorly functioning items, or to judge the appropriateness of the allowed time given the items in the test. But for a tightly constrained test with a known test population and previous data on test performance, as demonstrated in this article, the approach of using data from comparative judgements as a proxy for pre-testing assessments to estimate the relative difficulty of tests has potential where there may be security issues in exposing students or other interested stakeholders to assessment contents.

REFERENCES

Andrich, D. (1978a). A binomial latent trait model for the study of Likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology*, *31*, 84–98. doi:10.1111/j.2044-8317.1978.tb00575.x

Andrich, D. (1978b). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 451–460. doi:10.1177/014662167800200319

Andrich, D. (2015). The problem with the step metaphor for polytomous models for ordinal assessments. *Educational Measurement: Issues and Practice*, *34*, 8–14. doi:10.1111/emip.12074

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating item difficulty with comparative judgements. *ETS Research Report Series*, *2014*, 1–8. doi:10.1002/ets2.12042

Baird, J. (2007). Alternative conceptions of comparability. In P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 124–156). London, UK: QCA.

Baird, J., Cresswell, M., & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, *15*, 213–229. doi:10.1080/026715200402506

Bèguin, A. (2000). Robustness of equating high-stakes tests. PhD diss., The Netherlands: University of Twente.

Bejar, I. I. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, *7*, 303–310. doi:10.1177/014662168300700306

Bradley, R., & Terry, M. (1952). Rank analysis of incomplete block designs I: The method of paired comparisons. *Biometrika*, *39*, 324–345. doi:10.1093/biomet/39.3-4.324

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

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–94). London, UK: QCA.

Christensen, K., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*, 178–194. doi:10.1177/0146621616677520

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications Ltd.

DfE. (2010). *The importance of teaching: The schools white paper 2010*. London, UK: Department for Education.

DfE. (2013). *GCSE mathematics: Subject content and assessment objectives, DFE-00233-2013*. London, UK: Department for Education.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., Sireci, S. G., Swaminathan, H., Xing, D., & Rizavi, S. (2003). *Anchor-based methods for judgementally estimating item difficulty parameters (LSAC Research Report 98-05)*. Newtown, PA: Law School Admission Council.

Hanson, B. A., & Bèguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3–24. doi:10.1177/0146621602026001001

He, Q., & Wheadon, C. (2013). Using the dichotomous Rasch model to analyse polytomous items. *Journal of Applied Measurement*, *14*, 44–56.

He, Q., Anwyll, S., Glanville, M., & Opposs, D. (2014). An investigation of measurement invariance of the Key Stage 2 National Curriculum science sampling test in England. *Research Papers in Education*, *29*, 211–239. doi:10.1080/02671522.2012.742133.

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, *13*, 151–177. doi:10.1007/s10763-013-9497-6

Laming, D. (1990). The reliability of a certain university exam compared with the precision of absolute judgements. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *42*, 239–254. doi:10.1080/14640749008401220

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, 85–106. doi:10.1016/j.apsusc.2016.02.162

Linacre, J. (2017). *Winsteps® Rasch measurement computer program User's Guide*. Beaverton, OR: Winsteps.com.

Luce, R. D. (1959). *Individual choice behavior*. New York, NY: Wiley.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/BF02296272

Maydeu-Olivares, A. (2013). Goodness-of-Fit assessment of item response theory models. *Measurement: Interdisciplinary Research & Perspective*, *11*, 71–101. doi:10.1080/15366367.2013.831680

McGill, M. (2009). An investigation of unidimensional testing procedures under latent trait theory using principal component analysis. PhD Thesis, Virginia Polytechnic Institute and State University.

Melican, G. J., Mills, C. N., & Plake, B. S. (1989). Accuracy of item performance predictions based on the Nedelsky standard setting method. *Educational and Psychological Measurement*, *49*, 467–478. doi:10.1177/0013164489492020

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. doi:10.1177/014662169201600206

Ofqual. (2015). *A comparison of expected difficulty, actual difficulty and assessment of problem solving across GCSE maths sample assessment materials, Ofqual/15/5679*. Coventry, UK: The Office of Qualifications and Examinations Regulation.

Ofqual. (2017). Inter-board comparability of grade standards in GCSEs, AS, and A levels 2017. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/666765/Inter-board_comparability_of_grade_standards_in_GCSEs__AS_and_A_levels_2017_.pdf

Pae, H. K. (2012). A psychometric measurement model for adult English language learners: Pearson test of English academic. *Educational Research and Evaluation*, *18*, 211–229. doi:10.1080/13803611.2011.650921

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice*, *19*, 281–300. doi:10.1080/0969594X.2012.665354

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 166–206). London, UK: QCA.

Pyle, K., Jones, E., Williams, C., & Morrison, J. (2009). Investigation of the factors affecting the pre-test effect in national curriculum science assessment development in England. *Educational Research*, *51*, 269–282. doi:10.1080/00131880902892022

Rasch, G. (1960). *Probabilitistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Paedagogiske Institute.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer-Verlag.

Robinson, C. (2007). Awarding examination grades: Current progresses and their evolution. In P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 97–123). London, UK: QCA.

Shermis, M. D., & Chang, S.-H. (1997). The use of item response theory (IRT) to investigate the hierarchical nature of a college mathematics curriculum. *Educational and Psychological Measurement*, *57*, 450–458. doi:10.1177/0013164497057003006

Smith, E. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*, 205–231.

Steedle, J., & Grochowalski, J. (2017). The effect of stakes on accountability test scores and pass rates. *Educational Assessment*, *22*, 111–123. doi:10.1080/10627197.2017.1309276

Tan, J., & Yates, S. (2007). A Rasch analysis of the academic self-concept questionnaire. *International Education Journal*, *8*, 470–484.

Thurstone, L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273–286. doi:10.1037/h0070288

Wright, B., & Masters, G. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

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

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne, Australia: Educational Measurement Solutions.

Yen, S., Bené, N., & Huynh, H. (2000). *The effect of content integration on the construct validity of reading performance assessment*. Paper presented at the 2000 Annual Meeting of the National Council on Measurement in Education, New Orleans, LA, April 25–27.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x