

High Stakes Tests with Self-Selected Essay Questions: Addressing Issues of Fairness

Iasonas Lamprianou

*School of Education, University of Manchester
Manchester, UK*

This study investigates the effect of reporting the unadjusted raw scores in a high-stakes language exam when raters differ significantly in severity and self-selected questions differ significantly in difficulty. More sophisticated models, introducing meaningful facets and parameters, are successively used to investigate the characteristics of the dataset. The application of the Rasch models to the data showed that examinees could benefit significantly from being marked by lenient raters and by responding to less demanding essay questions. It was also shown that the third rater failed to adjust the raw scores in a way similar to the statistical adjustment by the Rasch models. The study discusses the consequences of reporting unadjusted raw scores with particular emphasis on issues of fairness.

Keywords: fairness, high-stakes testing, Rasch model, self-selected questions

Political pressure on schools for the accountability of pupil performance has resulted in the implementation of high-stakes assessment procedures in many educational systems around the world. These assessment procedures often take the form of formal written examinations.

Many consider the inclusion of free-response questions in large scale assessments a “mixed blessing” (Braun, 1988). Braun suggested that the use of essay questions was generally very useful because other types of abilities that could not be assessed by multiple-choice questions might be assessed by essay questions. However, he suggested that scoring problems might arise: a large number of raters (the terms raters and markers will hence be used interchangeably) had to be trained

and supervised, and the maintenance of consistent marking across raters, between raters, and across time was difficult to achieve.

According to Linacre and Wright (2002), four factors govern the scores assigned to the performance of an examinee. These are the ability of the examinee, the difficulty of the questions, the severity of the rater, and the way in which the rater applies the rating scale (the terms “rating scale” and “mark scheme” are used interchangeably in this article).

In practical settings, however, it has long been observed that a number of characteristics of the raters may intervene in the rating process (Guilford, 1954). For example, as early as the beginning of the 20th century, Thorndike (1920) discussed raters’ tendency to evaluate an examinee holistically as being generally good or inferior, without discriminating between different dimensions of behavior or performance (halo effect). Two years later, Kingsbury (1922) discussed situations in which raters consistently gave low or high ratings (severity–leniency) or hesitated to discriminate and tended to give similar ratings (range restriction).

Recent research on the differential severity of raters suggests that although they may sometimes exercise similar levels of leniency or severity (e.g., Rae & Hyland, 2001), it is frequently the case that they may disagree significantly in their ratings (Banerji, 1999; Barrett, 2001; Bonk & Ockey, 2003; Engelhard, 1992; Lumley, Lynch, & McNamara, 1994; Weigle, 1999). Moreover, it was suggested that raters not only varied in harshness generally, but could also exhibit differential harshness when scoring for different questions and for different sub-groups (Linacre, 1994).

A number of rater characteristics have been investigated to identify sources of differential rating style. Weigle (1998) reported that more experienced raters were more lenient than their less-experienced colleagues. Bonk and Ockey (2003) found that returning raters (i.e., more experienced) were more severe, but Barrett (2001) found mixed results.

Rater errors such as restriction of range, halo effect, and severity/leniency have been extensively investigated, but this has sometimes resulted in inconclusive or contradictory results. For example, Murphy, Jako, and Anhalt (1993) discussed the nature and the consequences of the halo effect and reviewed evidence that halo errors were not ubiquitous but inflated correlations among rating dimensions were not the norm. Murphy et al. asked for a moratorium on the use of halo effect indices as dependent measures in applied research.

Engelhard (1994) also reviewed several categories of rater errors basing his theoretical background on the work of Saal, Downey, and Lahey (1980) and Thorndike (1920). He agreed that the halo effect exists when a rater fails to distinguish between aspects of examinees’ responses in cases in which those should be conceptually distinct and independent. He described the central tendency effect as a self-restriction of the rater in awarding only a short interval of scores around the midpoint of the rating scale. Restriction of range, on the other hand, was described as a self-restriction of the rater in awarding a short range

of scores which could either be at the midpoint or anywhere else on the rating scale.

The stability of the harshness of a rater has also been investigated. Congdon and McQueen (2000) recently identified significant fluctuations in the harshness of raters across a marking period of seven days. Hoskens and Wilson (2001) also studied changes in raters' severity across time. They found that raters' severity tended to drift toward the mean when measured in five successive periods.

In the case when more than one rater marks a script (e.g., essay), a significant disagreement between the raters may result in marking supervisors, also referred to as table leaders, taking some action. Various discrepancy resolution methods have routinely been applied by different testing agencies, and Myford and Wolfe (2002) offer a comprehensive literature review of research that deals with the issue.

More than one methodology has been developed to deal with the issues of the differential severity of raters and of the differential difficulty of questions. Classical test theory (usually in the form of correlations and analysis of variance) served this area of research for many years. However, generalizability theory is also one of the major players in this area; it focuses on the dependability of measurements and it is widely used in reliability studies.

Longford (1994) was motivated by the generalizability theory and suggested that adjusting scores for differences in rater severity was not enough. He suggested that the reliability of the measurement of the harshness was also very important. The generalizability theory proved to be a useful tool in this area of research and has been widely used in the last decade (e.g., Bock, Brennan, & Muraki, 2002; Hurtz & Hertz, 1999; Rae & Hyland, 2001; Swartz et al., 1999). One of its major applications has been to identify and measure the rater effect, and in general, to measure the sources of measurement error and decide which of the sources are small enough to be ignored. A debate concerning the merits of the generalizability theory has been on for decades and the international literature has elaborated on the issue for some time (Marcoulides, 1993); however, the adjustment of individual examinee scores accounting for the differential harshness of the raters has not been one of the strong aspects of generalizability theory. For example, Linacre and Wright (2002) argued that the use of the many-faceted Rasch model has specific advantages over the generalizability theory. According to Linacre and Wright, the many-faceted Rasch model treats the examinees as individuals and attempts to free their ability measure from the distributional details of the other facets of the analysis (e.g., examinees, items, and rater).

This research investigates the validity of the uses of raw test scores in a national university entrance testing system. It indicates how the rank order of examinee abilities could change if appropriate statistical models were used to adjust for the differential severity of the raters and for the differential difficulty of the questions. The reluctance of policymakers to employ adjustments in order to achieve a fairer marking system is also discussed.

THE INSTRUMENT

The language test under study was part of a battery of tests of various subjects (e.g., mathematics, science, history). The test was administered in the summer of 1999 to secondary school graduates (18 year olds) in the context of a competitive high-stakes university entrance examination system (more examinees were seeking acceptance to universities than the universities would accept). The test was constructed to measure proficiency in the national language of a European country (the name of the country and the testing agency are treated as confidential).

The first part of the test consisted of six short-answer questions, which all examinees had to complete. These will be referred to as “core” questions one to six. The six core questions were designed to measure the “use of language” ability of the examinees. They asked the examinee to demonstrate knowledge of the meaning of words, to identify the root of words, to generate nouns from verbs, to use words properly to construct meaningful sentences, etc. The correct response to those questions was either a sentence (in those cases where the examinees were asked to generate, correct, or modify a sentence), or a single word (in the cases where they were asked to write a synonym). Questions one to four were worth two points. The examinees were awarded one mark as partial credit for a partially correct answer (e.g., in the case where the question asked them to give two synonyms but they were only able to give one correct synonym). Questions five and six were worth one mark. The raters followed strictly structured mark schemes for questions one to six.

The second part of the test consisted of four essay questions. Each examinee had to complete one of the four essay questions (hence, “self-selected” questions), which could be classified as a combination of narrative and persuasive writing. Although the specific essay prompts for the test cannot be disclosed, it may be said that the examinees were asked to explicate and elaborate on social issues, or discuss ideas and concepts (e.g., “The impact of illegal immigration on the political relationships between Europe and the Third World”). The examiners made an effort to formulate the four essay questions so that they would be equally difficult, but they only based their efforts on intuitiveness and experience, not on hard evidence or any formal theory.

The self-selected questions were marked on three criteria: Content, Language, and Structure. According to the examination board, the Language criterion focused on spelling, the richness of the vocabulary, and the clarity and style of writing.

The Content criterion focused on the appropriateness and originality of the ideas included in the essay and the adequacy of the arguments. The Content and Language criteria were somewhat interdependent because rich vocabulary and clarity of writing were very important to achieve a high Content mark given the limited number of words allowed (no more than 750 words).

Examinee Groups	Core questions						Self-selected questions (one to be completed)												
							Essay A			Essay B			Essay C			Essay D			
							Criteria												
	Q1	Q2	Q3	Q4	Q5	Q6	a	b	c	a	b	c	a	b	c	a	b	c	
A																			
B																			
C																			
D																			
Points	2	2	2	2	1	1	a=25, b=15, c=10 (Essay total: 50 points)												

Notes:

- white cells indicate non-administered questions (examinees completed one self-selected question);
- a, b, c indicate the three criteria which were exactly the same across the different themes (i.e., the richness of the content, the quality and appropriateness of the language - including the correctness of spelling - and the structure of the essay).

FIGURE 1
The structure of the data.

The Structure criterion focused on the appropriateness of the structure of the paragraphs and the structure of the essay in general (for example, the essay should include an introduction, main part, and conclusion).

Each examinee eventually received nine scores: one for each of the six core questions, and three for the self-selected question (one for each of the three criteria). Although detailed descriptions of “exemplary” material were available, the raters were left on their own to form a view of the quality of each essay, with respect to each criterion (i.e., up to 25 points for Content, 15 for Language, and 10 for Structure; see Figure 1). All in all, the examination agency considered the total score for the test to be a valid indicator of examinees’ “use of language” ability. The sum of the nine scores gave the raw score awarded by the rater to any examinee.

Although rating scales from 1–6 are usually used across rating dimensions to rate essays in examination settings, the agency that is responsible for the development of the specific test believes that it is possible to differentiate reliably between examinee performance using long scales (i.e., scale 0–25) for the Content criterion. This has been the tradition in the specific examination context for many

years, and the public, the teachers, and the test development organization seem to be satisfied with this arrangement.

For purposes of clarity, it is explained that the mark scheme supported the award of half marks in which a rater thought that he or she could rate the response of the examinee with such precision. For example, a rater might award 7.5 marks for a response that fulfilled the criteria for 7 marks, but was just failing the criteria for the 8 marks. Although this has been the focus of much discussion in recent years, both the raters and the test development organization felt that this practice was both practical and efficient.

The examinees were free to complete either the core questions or the essay first. After the marking procedure, the testing agency reports to the examinees their pass/fail classification. However, the actual rank order of the examinees is also very important once they pass the examination because this determines whether they will be awarded a place in a university of their first, second, third preference, and so on. The rank order of the examinees is also important to those who fail the initial pass/fail classification because additional places frequently become available prior to the start of the academic year.

THE RATERS

A total of 55 raters participated in the marking exercise. These were experienced, secondary education language teachers and most of them were returning raters (i.e., they had marked scripts at least once before). No information about their demographic characteristics was available. However, their first language needed to be the language of the test. They were selected from a pool of “high quality” teachers (suggested by language school inspectors), and they were paid a certain amount of money per script they marked. No official moderation (i.e., training) was carried out other than a short briefing. During the briefing, the raters gathered in a conference room and under the supervision of the senior test constructor they went through the scoring rubric. They discussed the suggested answer for each of the questions, and they had the opportunity to read and discuss exemplar responses on each of the questions. The session lasted for approximately four hours and at the end they were encouraged to ask for help from more experienced colleagues when needed.

AIMS

This study uses data from a high-stakes national test to investigate a number of issues:

- The effects of the differential difficulty of self-selected questions and the differential severity of raters on the test results when the raw scores are reported;

- The impact of the third rater on the test results and to compare this with the impact of other score adjustment models;
- How increasingly sophisticated models may be used to reveal the characteristics of the dataset by introducing meaningful additional facets and parameters.

THE DATASET

The dataset consisted of 5603 examinees (approximately 42% were male). They were all of the same nationality, they all had the language of the test as their first language, and they all came from government-supported schools, which followed the same National Curriculum and used the same textbooks. The sample included students from both rural and urban schools of various sizes (the number of students per school ranged from a few tens to a few hundreds). According to the testing agency, the sample is representative of the total population that took the examination as far as variables such as gender, geographical area, school type, age, and performance are concerned.

Approximately half of the examinees completed Essay “D”, one fifth completed Essay “A”, one fifth completed Essay “B,” and less than 5% of the examinees completed Essay “C.” Each of the scripts was marked by two raters who were selected randomly. The two markings were blind in the sense that the second rater was not aware of the first marking. In the case of a large discrepancy between the first two markings (more than 10% of the total score), the scripts were sent to a third rater. The third rater was aware of the scores awarded by the first two raters but did not know their identities or the identity of the examinee. If a third rater was not needed, the sum of the first two markings was reported. If a third rater was needed, the first two markings were discarded, and the third marking was doubled and used in their place. The third raters were acknowledged to be senior expert item-writers, as well as members of the committee who built the tests under consideration in this study. They were in charge of the fine-tuning of the scoring rubric and they were also responsible for the moderation/training of the raters.

Each rater marked all the questions on a given script in the order they appeared on the test. The scripts were randomly allocated to the raters in bunches of 50 scripts so all raters were ‘interconnected’ with at least one bunch of 50. For example, Rater 1 had 50 scripts in common with each one of the Raters 2, 3, 28, and 56, etc. All raters were therefore interconnected between them.

The testing agency instructed the raters that missing responses would be automatically marked as incorrect. The examinees had three hours to complete the test. The testing agency considered this time to be adequate and the test was not considered to be a speeded test.

METHOD

The structure of the data (see Figure 1) is similar to that of a standard test-equating task. The data collection design for this study takes the form of the Anchor-Test-Nonequivalent-Groups Design described by Petersen, Kolen, and Hoover (1993). According to this design, each of four different groups of examinees completes a different form of the same test (one of the four essays). Still, all groups complete a common test (the core questions), which is called “the anchor test.” The goal of test equating is to create sets of equivalent scores across the four forms. The results of an equipercentile test-equating for the four essays (each considered to be a different paper) using the core questions as an anchor paper is presented in the Appendix for purposes of comparison.

The use of the Rasch model is another option for this “test-equating” task and also has the advantage of adjusting scores for the differential harshness of the raters. The Rasch model is a probabilistic model used widely for purposes of educational measurement around the world. This model aims to describe the interaction between an examinee and a question, expressing the abilities of examinees and the difficulties of the questions onto a single linear scale (i.e., the logit scale). This serves the need for direct comparisons between the ability of a specific examinee and the difficulty of a specific question. Equation (1) illustrates the Rasch model in the case where an examinee n attempts to respond to question i which is scored on a scale from zero to k .

$$\log \left(\frac{P_{nik}}{P_{ni(k-1)}} \right) = B_n - D_i - F_k \quad (1)$$

where, P_{nik} is the probability of examinee n being assigned on question i , the score k , $P_{ni(k-1)}$ is the probability of examinee n being assigned on question i , the score $k - 1$, B_n indicates the ability of the examinee n , D_i indicates the difficulty of the question i , and F_k indicates the difficulty of score k in relation to score $k - 1$.

Each distinct score on a question (marked from 0 to k points) may be considered as a step. According to Equation (1), the interaction between an examinee’s ability, a question’s difficulty, and a step’s difficulty, define the score that an examinee is most likely to obtain on the question.

Equation (1) illustrates the case where all the questions on the test employ the same scale (e.g., 0 to 10 points), and it is assumed that the scale maintains the same meaning across the questions (e.g., a score of 5 has the same meaning for all questions). This is the Rating Scale model (Andrich, 1978; Wright & Masters, 1982) and it is the first model to be presented in this study.

In the case in which each question is modeled to show its own rating scale (e.g., when a score of 5 for one question is not equivalent to a score of 5 for another

question), then the equation becomes:

$$\log \left(\frac{P_{nik}}{P_{ni(k-1)}} \right) = B_n - D_i - F_{ik} \quad (2)$$

where F_{ik} indicates the difficulty of score k in relation to score $k - 1$ for question i . This is the Partial Credit model (Wright & Masters, 1982) and it is the second model to be presented in this study.

A more general version of the Rasch model is the many-faceted Rasch model (Linacre, 1994). This model is mathematically more complex than the two-faceted Partial Credit model (Equation (2)) but conceptually it is just a natural generalization that takes into consideration the harshness of the raters or any other facets that intervene in the procedure of measurement. A special case is the three-faceted model that includes the facets questions, examinees, and raters.

The three-faceted approach permits the calibration of a severity (or leniency) effect that may differentiate raters in their scoring. This approach allows removal of the influence of the differential severity of the raters from the score obtained by the examinee (Linacre, 1994). Raters could develop and use their own unique standards and may differ from the other raters in their leniency and severity (Lunz, Wright, & Linacre, 1990). As Lunz et al. (1990) explained “when judges vary in severity, raw scores are affected and decisions may be different” (p. 333). It is, therefore, important to use the three-faceted approach because it adjusts the scores for the differential severity of the raters and, in effect, lessens the worries of the examinees as to who marks their responses to the questions of the test.

Equation (3) demonstrates a three-faceted Partial Credit Rasch model.

$$\log \left(\frac{P_{nijk}}{P_{nij(k-1)}} \right) = B_n - D_i - C_j - F_{ik} \quad (3)$$

where, P_{nijk} is the probability of examinee n being assigned on question i , by the j th rater, the score k , $P_{nij(k-1)}$ is the probability of examinee n being assigned on question i , by the j th rater, the score $k - 1$, C_j is the harshness of rater j , and F_{ik} is the difficulty of being awarded score k instead of score $k - 1$ on question i .

The Selection of the Rasch Model

The Rasch model was considered to be more appropriate than other item response theory (IRT) models for this dataset for a number of reasons. First, the raw scores of the examinees were considered by the testing agency to be a sufficient statistic for the estimation of their underlying ability. In addition, the raters did not award more points for correct or partly correct responses to more difficult questions and did not penalize the examinees for incorrect or partly-correct responses to easier

questions. Finally, the nature of the open-ended questions did not encourage guessing. Overall, models that had weighted scores as sufficient statistics or incorporated pseudo-guessing parameters were not appropriate.

Two fit statistics were selected for this study: the INFIT Mean Square and the OUTFIT Mean Square (Wright & Stone, 1979), which tend to identify general aberrance in response patterns and do not focus on a specific type of misfit. This is an advantage because a fit statistic that focuses only on a specific type of aberrance may not have enough power to identify other types of aberrance (Klauser, 1995).

Much available literature encourages the use of the two statistics in the context of Rasch models (e.g., Smith, 1991, 2000; Wright & Masters, 1982; Wright & Mok, 2000). Both are approximately chi-square distributed (Wright & Mok, 2000), but no universal cut-off values have been accepted. Karabatsos (2000) suggested that the distributional properties of the two statistics could differ significantly across datasets. In the context of many-faceted Rasch measurement, various researchers have used slightly different ranges of acceptability for the fit statistics. For a number of studies (e.g., Engelhard, 1992, 1994; Lunz et al., 1990) the range of acceptance for examinee and rater fit was set at 0.6 to 1.5. The range of acceptance for the question fit was often set from 0.7 to 1.3. However, these are just rules of thumb and may not be appropriate for all datasets in every context.

This study does not depend overly on predefined cut-off values for the fit statistics based on totally arbitrary type-I error rates (e.g., 5%) because misfit and overfit should not be considered as a have/not have property. Elements in the facets of raters and questions with unusually high fit statistics (anything above or approaching 1.2) were not “dropped” as misfitting but were studied in more depth to identify the source of the aberrance. Examinees’ response patterns with fit statistics larger than 2 were also flagged for further investigation.

SPSS version 14 was used to run a factor analysis on the data as a preliminary investigation of unidimensionality. According to SPSS documentation, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy assesses whether the partial correlations between the test items are small. KMO values close to 1 are desirable, but according to rules of thumb, KMO values larger than 0.5 are generally encouraging to proceed with the factor analysis. In addition to the KMO statistic, Bartlett’s test of sphericity is used to test whether the correlation matrix is an identity matrix; in such a case, this might indicate that the factor model is inappropriate.

As it has been explained in the section where the instrument was discussed, the raters were free to award half marks in the case where they felt that they could reliably judge that a response fell between two score categories of the mark scheme. For example, a response might be awarded 7.5 marks because the rater felt that it was ‘better’ than the exemplar responses (discussed during the raters’ briefing) that were awarded 7 marks, but not as good as the responses that were awarded 8 marks. In order to run the Rasch analysis (only integers may be used

by the FACETS and Analysis Rasch packages, Linacre, 2006), all the scores were doubled. In effect, a scale with a range from 0–4 was turned to a scale from 0–8, in which the half marks were represented by the odd numbers. This does not affect the Rasch analysis, because the Rasch analysis does not imply that the scoring is originally done in integers; the Rasch analysis hypothesizes a hierarchy of ordered categories, no matter if they represent half marks, decimals or any other type of observations.

RESULTS

Although the maximum possible score on the test was 60 points, the total raw scores of the examinees were up to 120 points since each script was marked by two raters and their markings were summed. In the case of the scripts that were remarked by a third rater, the total maximum score is once again 120 points because the third marking was doubled and the first two markings were discarded.

Each of the 55 raters marked a slightly different number of scripts, but each of them marked an average of just over 200 scripts. Although the scripts were randomly allocated to each rater, a number of raters appeared to be awarding low points and others appeared to be awarding higher scores. The most “generous” rater awarded an average of 46.4 points and the least “generous” rater awarded an average of 25.2 points (Figure 2).

The mean total raw score for all the examinees was 72.3 ($SD = 18.4$). The smallest raw total score was 1 and the largest was 118, almost a perfect score. A few of the questions on the test appeared to be easy. Table 1 presents some statistics on the questions of the test.

Regarding the four essay topics, Essay A was completed by 1261 examinees and was the easiest (Content Mean = 15.9, Language Mean = 10.6, Structure Mean = 7.0). Essay B was completed by 1154 examinees and was relatively difficult (Content Mean = 12.8, Language Mean = 9.3, Structure Mean = 5.9 points). Essay C was completed by a very small number of examinees ($N = 179$) and was equally difficult (Content Mean = 12.6, Language Mean = 9.7, Structure Mean = 5.9 points). Essay D was of “moderate” difficulty (Content Mean = 14.6, Language Mean = 9.8, Structure Mean = 6.5 points) and was completed by 3009 examinees.

An analysis of variance (ANOVA) test found statistically significant differences between the average raw scores awarded to each essay ($F_{3,5599} = 115.164$, $p < 0.001$). The correlation between the total score on the core questions and the total score on the essay questions was 0.55 and was statistically significant ($p < 0.01$). In order to investigate the relationship between the four major components of the test (i.e., core questions, Content, Language, and Structure), a factor

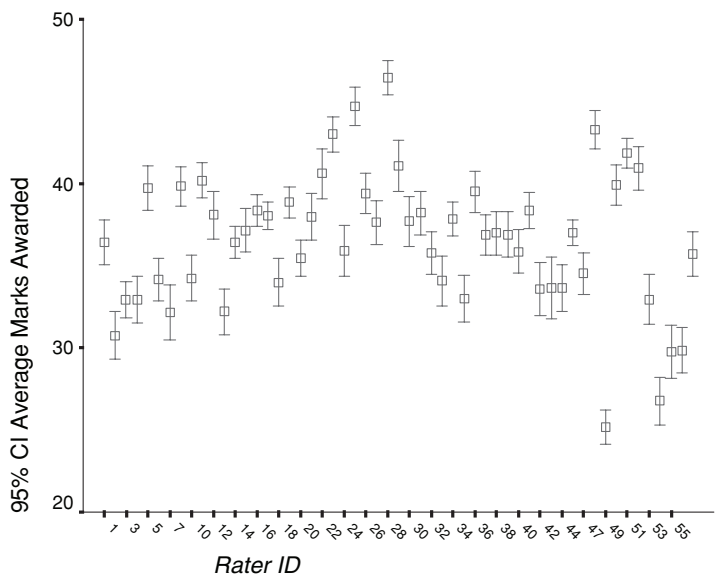


FIGURE 2
Average score awarded by each rater (with 95% confidence intervals).

analysis was run. Only one factor was extracted which explained approximately 68% of the variance. This result provided an initial indication that the test might be unidimensional. The KMO statistic was large (0.817) and Bartlett’s Test of Sphericity gave statistically significant results (approx. $\chi^2 = 29084.72$, $df = 6$, $p < 0.0001$).

A Two-Faceted Rasch Analysis

A two-faceted Rasch model was run in order to (a) investigate whether the data could form a unidimensional scale, and (b) determine whether the four essays (A to D) differed in difficulty; FACETS for Windows and Analysis (Linacre, 2006) software was used to run all Rasch models. No “rater effect” was investigated since no information about raters was used in the analysis; the two facets were the examinees and the questions. Only the data from the first two raters were used—the markings of the third rater were not included in the analysis. The purpose of this analysis was to first investigate the fit of the data to the Rasch model and then to determine whether the rank ordering of the examinees by raw scores was different from the rank ordering of examinees by their Rasch ability estimates.

TABLE 1
Available Points, Average Points Obtained, and Question-Total Correlations

	Core Questions						Overall for the Four Essays		
	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6	Content	Language	Structure
Available Points	2	2	2	2	1	1	25	15	10
Mean Points Achieved	1.05	0.90	1.55	0.85	0.80	0.35	14.55	9.90	6.50
% of Available Points Obtained	52.5%	45.0%	77.5%	42.5%	80.0%	35.0%	58.2%	66.0%	65.5%
Corrected Question-Total Correlation*	0.49	0.51	0.40	0.50	0.38	0.48	0.87	0.87	0.88

*The correlation between the question and the total score after subtracting score of the question from the total score.

The Rating Scale Model

Each of the criteria (Content, Language, and Structure) for the essay questions was modeled assuming that the raters used the marking scheme for that criterion in the same way across the four essays (e.g., 10 points awarded on Essay A for Content were assumed to be equivalent to 10 points awarded on Essay B for Content). In other words, three different rating scales were modeled: one for Content, one for Language, and one for Structure. This decision was in accordance with the intentions of the test developers who assumed that the examinees would be marked on each criterion across the four essays consistently using the same marking schemes (the same rating scale). The model used in this analysis is illustrated in Equation (1). Figure 3 shows the question hierarchy as well as the ability distribution of the examinees, and the rating scale for each of the three criteria and the core questions.

The fit of the data to the model was not satisfactory for all four criteria. The Content criterion for the four essay questions seemed to be generally misfitting; the INFIT mean-square (MNSQR) statistics for the Content criterion for the four essays were: 1.4, 1.8, 1.7, and 1.3 (see Table 2). This indicated that the points the examinees received on this criterion were less predictable than they should have been, given the scores these examinees received on the other two criteria. In other words, for the examinees with high INFIT MNSQR statistics, high scores on the Language and Structure criteria did not correspond to high scores on the Content criterion.

The mean ability of the examinees was 0.04 logits and the standard deviation was 0.49 logits. The distribution of the abilities resembled a normal distribution with the abilities ranging from -2.79 to 2.18 logits.

The average INFIT MNSQR statistic for the examinees was 1.20 ($SD = 1.02$). A number of examinees had high INFIT MNSQR statistics: 15.1% of the examinees had an infit MNSQR statistic larger than 2, 9.6% had an infit MNSQR statistic larger than 2.5, and 6.4% of the examinees had an infit MNSQR statistic larger than 3. This, in combination with the large INFIT MNSQR statistics for some of the questions, suggested that the fit of the data to the model was not satisfactory.

The rating scales of the three criteria had a number of problems identified by the analysis. For example, the rating scale of the Structure criterion had low frequencies on a number of scale scores; e.g., the score $1\frac{1}{2}$ was only observed 4 times, and score $2\frac{1}{2}$ was only observed 9 times. However, as it can be seen from Figure 4, the average Rasch ability of the 19 examinees who received $3\frac{1}{2}$ marks on this criterion is slightly larger than the average Rasch measure of the 863 examinees who received 4 marks. Although the difference is not practically important, it is indicative of the fact that raters who attempted awarding half marks may have been wrongly over-confident. This is a pattern that may be observed on all three essay criteria, although the relevant figures are not presented here for the sake of brevity.

TABLE 2
Question Statistics from the 2-Faceted Rating Scale and Partial Credit Models

<i>Question</i>	<i>2-Faceted Rating Scale Model</i>				<i>2-Faceted Partial Credit Model</i>			
	<i>Measure</i>	<i>S.E.</i>	<i>INFIT MNSQR</i>	<i>OUTFIT MNSQR</i>	<i>Measure</i>	<i>S.E.</i>	<i>INFIT MNSQR</i>	<i>OUTFIT MNSQR</i>
Question 1	−0.04	0.01	1.00	1.00	−0.03	0.01	1.00	1.00
Question 2	0.19	0.01	0.90	0.90	0.20	0.01	0.90	0.90
Question 3	−0.74	0.01	1.00	0.90	−0.74	0.01	1.00	0.90
Question 4	0.31	0.01	0.90	0.90	0.32	0.01	0.90	0.90
Question 5	0.68	0.02	0.90	0.90	0.69	0.02	0.90	0.90
Question 6	1.73	0.01	0.90	0.80	1.74	0.01	0.90	0.80
Essay A Content	−0.03	0.01	1.40	1.40	−0.11	0.01	1.40	1.40
Essay A Language	−0.23	0.01	1.10	1.10	−0.22	0.01	1.00	1.00
Essay A Structure	−0.23	0.01	0.90	0.90	−0.29	0.01	0.90	0.90
Essay B Content	−0.09	0.01	1.80	1.80	−0.05	0.01	1.50	1.50
Essay B Language	−0.31	0.01	1.30	1.30	−0.27	0.01	1.10	1.10
Essay B Structure	−0.28	0.01	1.00	1.00	−0.26	0.01	0.90	0.80
Essay C Content	0.12	0.01	1.70	1.70	0.07	0.01	1.30	1.40
Essay C Language	−0.17	0.02	1.30	1.20	−0.08	0.02	1.00	1.00
Essay C Structure	−0.09	0.02	1.30	1.30	−0.06	0.02	1.20	1.20
Essay D Content	−0.16	0.01	1.30	1.30	−0.17	0.01	1.40	1.40
Essay D Language	−0.32	0.01	1.00	1.00	−0.36	0.01	1.10	1.10
Essay D Structure	−0.34	0.01	0.80	0.80	−0.37	0.01	0.90	0.90
Mean	0.00	0.01	1.10	1.10	0.00	0.01	1.10	1.10
SD	0.51	0.01	0.30	0.30	0.52	0.01	0.20	0.20

[illegible]

Key

S1 to S6: The six core questions

S7 to S9: Content, Language and Structure criteria (the same scale for all essays)

FIGURE 3
The question hierarchy and examinee ability distribution (2-faceted rating scale).

The Partial Credit Model

In order to investigate whether the raters used the rating scale differently with each essay, a two-faceted Partial Credit Rasch analysis was run (i.e., each criterion for each essay was modeled to have its own scale).

DATA				QUALITY CONTROL			STEP	
Category	Counts	Cum.	Avg	Exp	OUTFIT	CALIBRATIONS		
Score	Used	%	Meas	Meas	MnSq	Measure	S.E.	
0	54	0%	0%	-.83	-.83	1.1		
1/2								
1	41	0%	1%	-.71	-.61	.7	-.43	.15
1 1/2	4	0%	1%	-.41	-.49	1.1	1.78	.12
2	234	2%	3%	-.44*	-.41	.8	-4.52	.11
2 1/2	9	0%	3%	-.29	-.33	1.2	2.89	.06
3	476	4%	7%	-.28	-.25	.8	-4.26	.06
3 1/2	19	0%	7%	-.11	-.18	1.0	3.00	.04
4	863	8%	15%	-.14*	-.11	.8	-3.96	.04
4 1/2	53	0%	16%	-.04	-.04	.8	2.72	.03
5	1222	11%	27%	.03	.04	.9	-3.14	.03
5 1/2	91	1%	27%	.18	.11	1.1	2.67	.03
6	1761	16%	43%	.17*	.19	.8	-2.81	.03
6 1/2	193	2%	45%	.28	.28	.8	2.45	.02
7	2172	19%	64%	.35	.37	.8	-2.10	.02
7 1/2	167	1%	66%	.48	.46	.8	2.98	.02
8	2381	21%	87%	.58	.56	.9	-2.15	.02
8 1/2	162	1%	88%	.75	.66	.7	3.30	.03
9	891	8%	96%	.81	.76	.9	-1.00	.03
9 1/2	28	0%	97%	1.12	.87	.6	4.27	.05
10	385	3%	100%	.97*	.98	1.0	-1.70	.06

Key

1st column (Score): The mark categories of the rating scale (including the half marks that the raters were allowed to award when they felt that they could do so reliably)

2nd column (Used): The frequency by which the mark category was observed

3rd column (Counts %): The percentage that corresponds to the frequency of the 2nd column

4th column (Cum %): The cumulative percentage of the 3rd column.

5th column (Avg. Meas.): The average Rasch ability measure of the examinees who were awarded that mark category

6th column (Exp. Meas.): The expected average Rasch ability measure of the examinees who were awarded that mark category

7th column (OUTFIT MnSq): The OUTFIT mean square statistic of the mark category

8th column (Measure): The Rasch difficulty estimate of the mark category

9th column (S.E.): The standard error of the difficulty estimate of the 8th column

FIGURE 4

The rating scale for the structure criterion of Essay B (2-faceted rating scale).

The results of the analysis indicated that the overall fit of the data to the model improved significantly compared to the more restricted model that was previously presented. Table 2 shows that the largest INFIT MNSQR statistic for the questions was only 1.5, compared to that of 1.8 for the previous model. The second largest INFIT MNSQR statistic was just 1.4, compared to 1.7 for the previous model. Although the INFIT MNSQR statistics for the questions were reduced significantly,

the Content criterion still had somewhat larger INFIT MNSQR statistics; the misfit for the other criteria decreased substantially.

The mean ability of the examinees was 0.02 logits and the standard deviation was 0.46 logits. The distribution of the abilities resembled a normal distribution with the abilities ranging from -4.3 to 2.3 logits. The fit statistics for a few of the examinees improved. The average INFIT MNSQR was 1.19 ($SD = 0.98$): 14.7% of the examinees had an INFIT MNSQR statistic larger than 2; 9% of the examinees had an INFIT MNSQR statistic larger than 2.5; and 6% of the examinees had an INFIT MNSQR statistic larger than 3. Overall, the response patterns of the examinees, as well as the table of residuals, were investigated thoroughly in order to evaluate the model data fit, and to visualize more effectively the meaning of the larger INFIT MNSQRs. The results were very indicative: in most of the cases, most of the misfit was due to only one highly unexpected score. For example, examinee 4648 had one of the worst INFIT MNSQRs compared to the other examinees (INFIT MNSQR = 3.3) due to one highly unexpected score on the Content criterion of Essay A. Examinee 4648 received a score of 30 points, whereas the expected score was 39.2. In another case, examinee 2465 (INFIT MNSQR = 3.1) had a single very unexpected score of 32 points on the Content criterion of Essay A whereas the expected score was 40.2 points.

TABLE 3
Question Statistics from the Three-Faceted Partial Credit Analysis

<i>Question</i>	<i>Measure</i>	<i>S.E.</i>	<i>INFIT MNSQR</i>	<i>OUTFIT MNSQR</i>
Question 1	-0.02	0.01	1.1	1.1
Question 2	0.23	0.01	1.0	1.0
Question 3	-0.79	0.01	1.0	1.0
Question 4	0.35	0.01	1.0	1.0
Question 5	0.73	0.02	1.0	0.9
Question 6	1.83	0.01	0.9	0.9
Essay A Content	-0.11	0.01	1.1	1.1
Essay A Language	-0.24	0.01	0.9	0.9
Essay A Structure	-0.32	0.01	0.8	0.8
Essay B Content	-0.04	0.01	1.4	1.4
Essay B Language	-0.29	0.01	1.1	1.1
Essay B Structure	-0.28	0.01	0.8	0.8
Essay C Content	0.09	0.02	1.4	1.4
Essay C Language	-0.08	0.02	1.0	1.0
Essay C Structure	-0.06	0.02	1.2	1.2
Essay D Content	-0.19	0.01	1.2	1.2
Essay D Language	-0.40	0.01	1.0	1.0
Essay D Structure	-0.42	0.01	0.8	0.8
Mean	0.00	0.01	1.0	1.0
SD	0.55	0.01	0.2	0.2

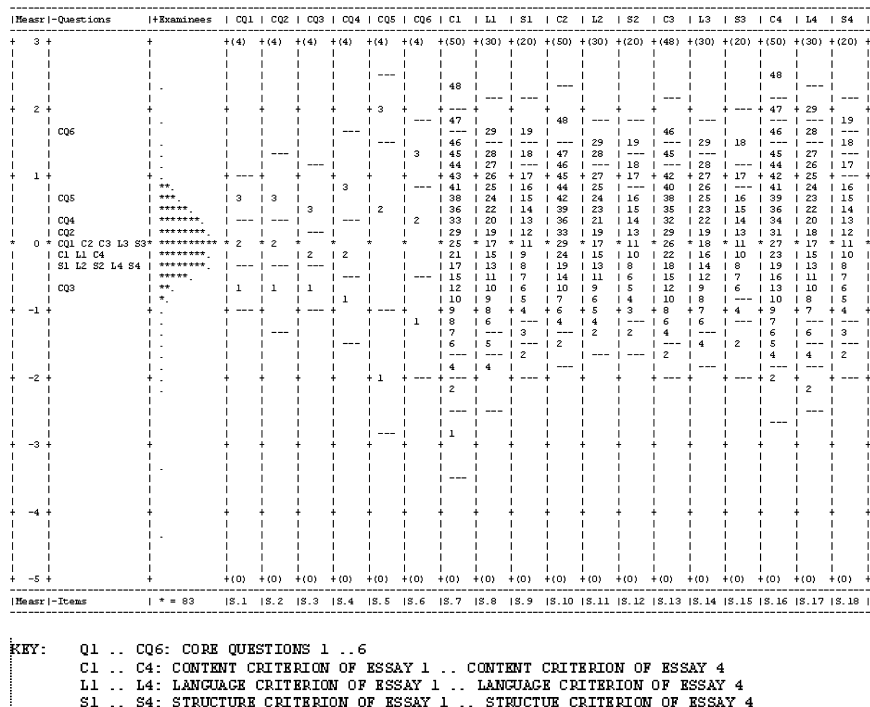


FIGURE 5

The question hierarchy and examinee ability distribution (2-faceted partial credit).

The results indicated that the criteria were differentially difficult across the four essay questions, which suggests that the raters did not use the rating scale for a given criterion in precisely the same way when marking examinees' responses to the four essays (Table 3). For example, the Content criterion for Essay C was 0.25 logits more difficult than that of Essay D. This was equal to 0.54 standard deviation units of examinee ability. The Language criterion for Essay C was 0.28 logits more difficult than that of Essay D (0.61 standard deviation units of examinee ability) and the Structure criterion for Essay C was 0.31 logits more difficult than that of Essay D (0.67 standard deviation units of examinee ability). In all of the above cases the differences were statistically significant at the 0.01 level.

Figure 5 illustrates the question hierarchy of the model, as well as the distribution of the ability of the examinee and the steps thresholds.

The examinees were ranked according to their raw scores and they were re-ranked according to their estimated ability measure. The lack of correspondence between the raw score percentiles and the Rasch ability measure percentiles show the bias due to the use of the unadjusted raw scores. The examinees' percentiles

could change considerably if their raw scores were adjusted to take into account the differential difficulty of the essay questions. The examinees who completed Essay A increased their percentile by approximately 3.9 centiles, those who completed Essay B showed practically no change and those completed Essay D decreased their percentiles by 2.3 centiles. However, the examinees who completed Essay C increased their percentiles by 10.8 centiles.

The Three-Faceted Rasch Model

The need for a three-faceted model resulted from the realization that the two raters who marked each script often disagreed in their markings. The total scores of the second rater explained a relatively small percentage of the variance of the scores of the first rater ($r = 0.65$, $R^2 = 0.42$). On average, the two raters who marked each script differed by 5.7 points ($SD = 5.3$). The minimum difference between the two raters was zero, and the maximum difference was 45 (out of a maximum possible of 60 points for the whole test). In 20% of the cases the markings of the two raters differed by more than 11 points.

Leaving the core questions aside, the difference between the essay total scores awarded by the two raters ranged from 0 to 45 points (out of 50 possible points for the three criteria), with an average absolute difference of 6.5 points ($SD = 5.2$). In order to accommodate for the differential harshness or leniency of the raters, as well as for the differential difficulty of the essay questions, a three-faceted Partial Credit Rasch model was run. The facets modeled were the examinees, the raters, and the questions. Each criterion had its own scale definition (see Equation (3)).

The average ability of the examinees was 0.04 logits ($SD = 0.50$). The average question difficulty was 0 ($SD = 0.55$). The average raters' harshness was 0 ($SD = 0.22$). The fit of the examinees improved considerably. The average examinee INFIT MNSQR statistic was 1.06 ($SD = 0.81$) compared to 1.19 and 1.2 from the two-faceted Partial Credit and Rating Scale analysis respectively. In addition, the percentage of examinees with INFIT MNSQR statistics larger than 2 was 10%, the percentage of examinees with INFIT MNSQR statistics larger than 2.5 was 5.6%, and the percentage of examinees with INFIT MNSQR statistics larger than 3 was 4.1%.

Similarly satisfactory was the fit of the majority of the raters with the exception of a few. For example, rater 7, who was one of the harsh raters with a severity estimate of 0.24 logits (std. error 0.01), had the largest fit statistic (an INFIT MNSQR statistic of 2.0). After a more in-depth investigation it was revealed that the source of misfit for this rater was a relatively small number of unexpected markings to examinees with high ability. Other raters had smaller INFIT MNSQR statistics (e.g., 3 raters had INFIT MNSQR statistics near 1.6). The rest of the raters had smaller INFIT MNSQR statistics. The average INFIT MNSQR value of the raters was 1.1 ($SD = 0.3$).

The question INFIT MNSQR statistics (Table 3) are smaller than the comparable statistics from the two-faceted Partial Credit model. The INFIT MNSQR statistics were all within an acceptance range of 0.6 to 1.2, except for two essay criteria which were the Content criteria for essays B and C). These two essays had the largest discrepancies between the markings of the two raters. It is likely that agreement between the raters and consistent application of the mark scheme was very difficult to achieve for the two essays. The average INFIT MNSQR value for the questions was 1.0 ($SD = 0.2$).

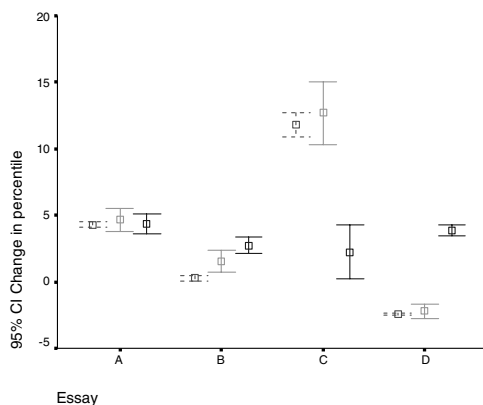
The model-data fit statistics relating to the functioning of the categories of the scales were generally satisfactory. Only a few categories of the Content criteria of the four essay questions had large fit statistics but in most of the cases this could simply be attributed to chance. For example, the OUTFIT MNSQR of the score 19 points on the Content criterion of Essay B was 3.7. However, after a closer inspection, it was found that only three examinees were awarded this category and their average ability happened to be unexpectedly high; e.g., their average ability was too high (-0.16 logits) compared to the average ability of the 186 examinees who scored 20 points and had an average ability -0.32 logits.

The results also indicated that the criteria for each of the four essay questions were differentially difficult (Table 3), which was in agreement with the results of the two-faceted Partial Credit model.

The raters were found to vary considerably in harshness. Their harshness estimates ranged from -0.58 to 0.59 logits. In contrast, the examinee ability estimates ranged from -1.80 to 2.24 logits. Taking into consideration the fact that the standard deviation of the examinee ability distribution was only 0.50 , the difference between the harshest and the most lenient rater could benefit or disadvantage an examinee by exactly two standard deviation units of ability.

The scores of the examinees who were marked by harsh raters were adjusted upward and as a result their ranks increased (e.g., one examinee moved from the 29th to the 70th percentile after adjustment for the harshness of the rater and the difficulty of the question). On the other hand, the score of examinees who had their scripts marked by very lenient raters were adjusted downwards and therefore, their ranks decreased. On average, the scripts marked by the harshest rater (rater 49) increased by approximately 18 centiles. The scripts marked by the most lenient rater (rater 25) decreased by approximately 14 centiles (Figure 6).

The application of the rating scale had very similar problems to those described previously for the 2-faceted Rating Scale model. Figure 7 shows rating scale statistics for the Structure criterion of Essay B. It may be suggested from the statistics that the raters who attempted awarding half marks did so at the expense of the reliability of marking, in the sense that the average Rasch ability of the examinees who received the half marks does not increase monotonically with the scale categories. Moreover, the low frequency of the half-mark scores lead to high errors of measurement, making the difficulty estimates of the rating scale rather unstable.



Key:

The dotted line presents the change in the percentile of examinee ability when the two-faceted model is used instead of the raw score awarded by the first two raters;
 The second line (faded) presents the change in the percentile of examinee ability when the three-faceted model is used instead of the raw score awarded by the first two raters;
 The third line (darker) presents the change in the percentile of examinee ability when the marking of the third rater is used instead of the raw score awarded by the first two raters.

FIGURE 6

Average percentile change and confidence intervals from the use of the adjusted scores.

The Third Rater

For 2408 examinees (i.e., 43% of the sample) a third rater was needed because the scores of the first two raters differed substantially. The exact agreement between the first two raters was 54% (exact agreement on 54% of the observed scores) whereas, probabilistically speaking, the Rasch model would expect only 29.8% agreement, after taking into account the differential harshness of the raters. All in all, FACETS (Linacre, 2006) computed that there were 50,427 marking opportunities (the number of examinees by the number of items in the test), and exact agreement was achieved in 27,220 cases.

The third marking was introduced to resolve the disagreement between the first two raters. The third rater did not have any statistical information about the script e.g., the rater did not know that a Rasch analysis would compensate the script for being marked by two severe raters. The identity of the candidate who produced the script was not revealed, and the third rater did not have any information about the performance of the candidate on other tests.

However, the dataset did not include the identity of the third rater. For the purposes of this analysis we will assume that the raters who acted as “third

DATA				QUALITY CONTROL			STEP	
Category	Counts	Cum.		Avge	Exp.	OUTFIT	CALIBRATIONS	
Score	Used	%	%	Meas	Meas	MnSq	Measure	S.E.
0	27	1%	1%	-.86	-.88	1.2		
1/2								
1	23	1%	2%	-.77	-.73	.8	-.64	.22
1 1/2	1	0%	2%	-.44	-.61	1.2	2.47	.17
2	85	4%	6%	-.52*	-.51	1.0	-5.00	.17
2 1/2	3	0%	6%	-.41	-.43	1.1	2.87	.11
3	153	7%	13%	-.36	-.35	.9	-4.32	.11
3 1/2	5	0%	13%	-.21	-.27	.6	3.11	.08
4	255	11%	24%	-.25*	-.19	.7	-4.16	.08
4 1/2	18	1%	25%	-.11	-.11	.8	2.50	.06
5	284	12%	37%	-.04	-.03	.7	-2.83	.06
5 1/2	20	1%	38%	.04	.05	.5	2.66	.06
6	396	17%	55%	.12	.14	.8	-2.89	.06
6 1/2	29	1%	56%	.26	.23	.7	2.80	.06
7	409	18%	74%	.32	.32	.7	-2.37	.06
7 1/2	22	1%	75%	.40	.42	.9	3.30	.06
8	379	16%	91%	.55	.53	.8	-2.37	.06
8 1/2	21	1%	92%	.64	.64	.8	3.48	.09
9	120	5%	97%	.88	.75	.7	-1.05	.09
9 1/2	2	0%	98%	.86*	.85	.9	4.89	.15
10	56	2%	100%	.95	.96	1.1	-2.43	.15

Key

1st column (Score): The mark categories of the rating scale (including the half marks that the raters were allowed to award when they felt that they could do so reliably)

2nd column (Used): The frequency by which the mark category was observed

3rd column (Counts %): The percentage that corresponds to the frequency of the 2nd column

4th column (Cum %): The cumulative percentage of the 3rd column.

5th column (Avge. Meas.): The average Rasch ability measure of the examinees that were awarded that mark category

6th column (Exp. Meas.): The expected average Rasch ability measure of the examinees that were awarded that mark category

7th column (OUTFIT MnSq): The OUTFIT mean square statistic of the mark category

8th column (Measure): The Rasch difficulty estimate of the mark category

9th column (S.E.): The standard error of the difficulty estimate of the 8th column

FIGURE 7

The rating scale for the structure criterion (3-faceted partial credit).

raters” worked in a perfectly consistent way—this was also the assumption of the testing service, based on the high expertise and experience of the third raters. All third markings will be treated as if they have come from a single “third rater.”

The average absolute difference between the score awarded by the third rater and the sum of the two blind markings was just below 3.6 points (out of 120

possible points), and the standard deviation was 3.2 points. The median difference was 3 points. Just below 10% of the examinees had their score changed by 7.5 points or more when a third rater marked their scripts. The third rater did not provide a detailed score (broken down by item), therefore it was not possible to compute an overall agreement for each rating opportunity as was done for the first and second raters. The exact agreement between the first, the second and the third raters on a scale from 0 to 120 was just below 2%. However, this would be expected since the scale was too large to expect exact agreements between the raters.

Figure 6 illustrates the magnitude of the “correction” to the percentile of the examinees’ ability, caused by implementing each of the three: (a) the two-faceted model, (b) the three-faceted model, and (c) the third rater.

Figure 4 indicates that introducing the third rater did not affect examinees’ rankings in the same way that analyzing the data with the two Rasch models did. For example, the average change to the percentile of the examinees resulting from the inclusion of the third raters’ markings was statistically the same across the four essays. However, both the three-faceted and the two-faceted models rewarded the examinees who completed Essays A, B, and C (harder essays), and penalized the examinees who completed Essay D (the easiest essay). In a sense, these results are not surprising, because the Rasch model took into consideration the differential difficulty of the essays, and therefore, adjusted the scores of the examinees accordingly; however, the third rater was asked to resolve the discrepancy between the two raters but may have failed to take into account the fact that each essay might be of different difficulty. It is reasonable that under such circumstances the third rater would not be able to compensate the examinees for answering a more difficult question.

A Comparison between the Four Methods

The correlation between the raw scores of the examinees (the scores awarded by the first two raters) and the adjusted scores by each of the four methods is presented on Table 4. The results suggest that the rank order of the examinees does not change substantially when different adjustment methods are used; the correlations are very high and range from 0.91 to 0.99. However, this is largely misleading since in the context of pass/fail high stakes and competitive examinations, we are mostly interested in the effect of the four adjustment methods on the scores of the high achieving examinees. This is because the pass/fail decision is very important but once this is established, the rank order of the examinees is used to determine who would enter the university of first choice, second choice, and so on.

In order to measure the effect of each one of the adjustment methods on the high achieving examinees, three groups of examinees were identified using their

unadjusted raw score: those at the 20%, 10%, and 5% top achieving groups. What was then investigated was how many of those examinees would remain in the top achieving groups after their scores were adjusted using each of the three methods. Only 82.5% of the examinees of the 20% top achieving group remained in the group after the three-faceted model was used. The percentages for the two-faceted model and the third rater method were 94.6% and 90.9%, respectively. Approximately 79.5% of the examinees of the 10% top achieving group remained in the group after the three-faceted model was used. The percentages for the two-faceted model and the third rater method were 97.6% and 90.4% respectively. Finally, approximately 69.1% of the examinees of the 5% top achieving group remained in the group after the three-faceted model was used. The percentages for the two-faceted model and the third rater method were 89.2% and 91.4% respectively.

The large effect of the three-faceted model (especially compared to the very small effect of the third rater method) on the high-achieving group (approximately 69.1% of the examinees of the 5% top achieving group remained in the group) is not only due to the logarithmic function which tends to be linear in the middle of the distribution but steeper at the tails (see Figure 8), but also due to the fact that harshness of the first two raters was also taken into account. This is probably an area where the third rater method has no effect: by design, the third rater could now know whether the first two raters were harsh or lenient.

TABLE 4
Correlations between the Candidate Scores Before and After the Adjustments

	<i>2-faceted Rating Scale model</i>	<i>2-faceted Partial Credit model</i>	<i>Many-faceted Rasch model</i>	<i>Third rater</i>
Score from the first two raters				
Pearson Correlation	0.97	0.96	0.93	0.99
Sig. (2-tailed)	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
N	5603	5603	5603	2471
2-faceted Rating Scale model				
Pearson Correlation		1.00	0.95	0.95
Sig. (2-tailed)		$p < 0.001$	$p < 0.001$	$p < 0.001$
N		5603	5603	2471
2-faceted Partial Credit model				
Pearson Correlation			0.95	0.95
Sig. (2-tailed)			$p < 0.001$	$p < 0.001$
N			5603	2471
Many-faceted Rasch model				
Pearson Correlation				0.91
Sig. (2-tailed)				$p < 0.001$
N				2471

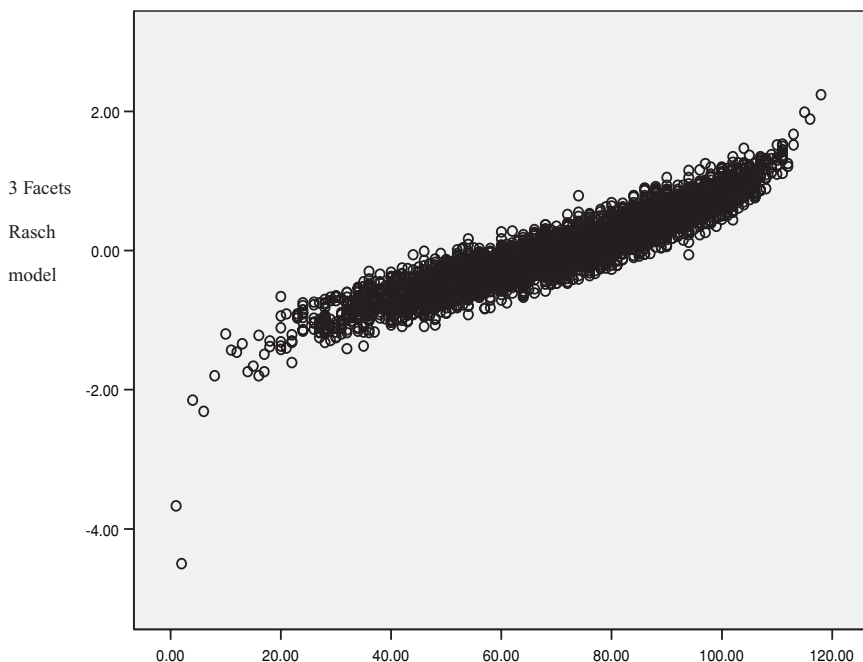


FIGURE 8

The relationship between the 3-facets adjusted scores and the sum of the raw scores awarded by the first two raters.

The Effect of the Adjustment on the Pass/Fail Classification

The passing rate was determined by the capacity of the Universities to accept first-year students. From the sample of 5603 examinees who took the test, 2645 of them were awarded a place to further their education.

All three methods (the two-faceted model, the three-faceted model, and the third rater) caused some changes in the raw score pass/fail classification of the examinees. Overall, the two-faceted Partial Credit model had the smallest impact on the classification. When it was applied, 3.2% of the examinees changed their status from Pass to Fail, and 2.8% of the examinees changed their status from Fail to Pass. The largest impact was caused by the three-faceted Rasch model. When it was applied, 12% of the examinees changed their status from Pass to Fail, and 10.7% of the examinees changed their status from Fail to Pass. The percentages for the third rater were 3.9% and 3.5%, respectively.

Overall, the classifications of 634 examinees were affected when the three-faceted Rasch model was applied to the data. A large percentage of the examinees

marked by the severe raters passed the test whereas the original classification based on the raw score failed them. For example, approximately 28% of the examinees marked by rater 49 (the severest rater) passed when the three-faceted model was used, but they failed based on the unadjusted raw scores. None of the examinees marked by this rater passed with their raw scores but failed when the three-faceted model was used.

On the other hand, a large percentage of the examinees who were marked by the most lenient raters failed when the three-faceted model was used, but they passed when their raw score was used. For example, more than 18% of the examinees marked by rater 25 actually failed when the three-faceted model was used, but they passed when their raw scores were used. None of the examinees marked by rater 25 passed according to the three-faceted model and failed when the raw score was used.

Finally, the effects of the three-faceted Rasch model are very different from the effects of the third rater on the pass/fail classification. When the effects of the two methods are compared, 312 examinees who were failed by the third rater actually passed the examination when the three-faceted model was used. Also, 312 examinees who passed the examination when the ratings of the third rater were used, actually failed when the three-faceted model was used.

DISCUSSION

In the context of this study, the policymakers of the testing agency responsible for the examinations made a number of assumptions. First, they assumed that the self-selected essay questions would be of approximately the same difficulty. This was not supported by the results of the analysis. The four essay questions were significantly different in difficulty, although it was not clear what exactly could have been the source of the differential difficulty. One side effect of this observation may be the fact that the third rater failed to compensate the examinees who attempted the most difficult essays. On the other hand, the Rasch analysis not only compensated the examinees who attempted the most difficult essays, but also penalized those who attempted the easiest ones. As a result of this, the adjustment of the third rater was not similar to the adjustment made by the Rasch model.

The results of this study are in agreement with previous research that indicates that questions designed to be of the same difficulty do not always function in that way. For example, researchers have found that even experts cannot predict the difficulty of an essay question before the actual administration of the test. Hamp-Lyons and Mathias (1994) asked experts to comment on the difficulty of essay questions. The results of the study showed that expert judges showed substantial agreement on issues such as prompt difficulty, prompt task type, and difficulty of prompt task type. "However, the patterns shown by the score data ran

in a direction which was the reverse of that predicted by the “expert” judgments” (Hamp-Lyons & Mathias, 1994, p. 49).

The issue of the differential difficulty of self-selected questions has raised much discussion in the past, not only in the area of school-leaving or university entrance exams, but also in the area of professional examinations. For example, Schlesinger (1995) argued that the fairness of the examination may be jeopardized:

The primary concern is ensuring a fair examination . . . since some of the items are bound to be easier to work than others, a candidate choosing the simplest . . . items would have a better chance of passing than another candidate making different selections. (p. 63)

It is also possible that more able examinees are avoiding the more challenging questions and answering the easy ones; this however could further disadvantage the less able examinees. This could also create problems while estimating the difficulty of each of the self-selected questions.

The New South Wales Board of Studies (2002) has put the issue in a different perspective arguing that the major issue is not so much whether self-selected questions are indeed equally difficult, but rather whether students choosing different options are treated fairly in the marking if the raters know that those options are not of the same difficulty. All these issues would probably need to be taken into account in a context where examinees have the liberty to choose which questions to respond to. How this could be taken into account statistically—if at all feasible—and how the human-adjustment methods (e.g., third rater) could deal with those issues is a different story.

Second, it was assumed that the raters would apply the marking scheme in the same way consistently across the four essays. However, when each essay was modeled to reveal the scale structure of the rating scale as raters applied it to each criterion, then the fit of the data to the model improved. This is a strong indication that the same raw score did not imply the same quality of work across the essays (e.g., a raw score of 10 out of 25 on Content did not imply the same quality of content across the four essays). To the degree that this is related to the training of the raters, the results of the analysis may suggest that the briefing the raters had in this particular case may not have been adequate in order to enable them to apply the rating scale consistently across the four essays. Another equally plausible interpretation of the results might be that the marking scheme was not prepared carefully enough to enable the raters to apply it consistently across the four essays.

Third, it was assumed that all the raters would mark interchangeably (i.e., they would exercise approximately the same severity) but this was not supported by the analysis. The range of the raters’ harshness was two times the standard deviation of the distribution of examinees’ ability when the three-faceted model was used.

It was also shown that the use of a model that would take into consideration the differential harshness of the raters would cause significant changes in the rank order of the examinees' performance. These results are in agreement with previous research which has shown that raters may differ significantly in the level of severity they exercise (Banerji, 1999; Bonk & Ockey, 2003; Engelhard, 1992; Linacre, 1994; Lumley et al., 1994; Weigle, 1998, 1999).

Fourth, it was assumed that the introduction of the third rater would manage to smooth any discrepancies between the first two raters. This was successful only to a degree. The use of a third rater in score resolution procedures has been suggested by previous research. For example, Johnson, Penny, and Gordon (1999) argued that the use of a third rater as a discrepancy resolution technique is not without merit. The results of this study lend support for their position that the third rater actually managed to smooth some of the unfairness introduced by raters who exercise extreme severity/leniency.

Johnson, Penny, and Gordon (2000), however, indicated that caution is needed when deciding how and when third rater resolution techniques are used. In any case, other recent research concluded that "implementing a discrepancy resolution procedure is not sufficient in and of itself for quality control monitoring" (Myford & Wolfe, 2002, p. 319). Myford and Wolfe found that the use of methods like the multi-faceted Rasch model may additionally be used to identify atypical ratings that would be missed by other resolution techniques. This study agrees that the use of the three-faceted model, in the specific context of the study, was a preferred adjustment method over the use of the third rater. However, it is important to keep in mind the fact that a three-faceted Rasch model takes into account additional factors when adjusting the scores of the examinees; the other models, i.e., the two-faceted Rasch model or the third rater, are blind to the differential harshness of the raters. A shift from one adjustment method to the other may be seen as a political decision, and a shift towards a three-faceted Rasch model would highlight that the testing service acknowledges explicitly (for the first time in this specific situation) that it is not feasible to train its raters so that they would mark in a practically similar way—thus the need for the three-faceted Rasch model adjustment.

Having said the above, another important issue that needs to be discussed—in relation to the above—is that of the face validity of the test results. In contrast to theoretical academic research, the discrepancy resolution techniques are partly used in order to enhance the face validity of high-stakes examinations. Parents, students, and other interested parties feel reassured when they know that a senior rater will have the opportunity to review their script in the case when the two first raters disagree. Depending on the expertise and seniority of the third rater, it is possible to defend any of the usual discrepancy resolution techniques such as replacing the first two scores with the score awarded by the third rater, or averaging the scores of all three raters, etc. (these have been widely discussed by Johnson, Penny, Fisher, & Kuhs, 2003; Johnson, Penny, & Gordon, 2001). The use of a statistical technique,

such as the Rasch model, may not have the same effect since it is not likely that all stakeholders will be able to understand such a complex technique.

Critiques of statistical resolution techniques might also argue that human raters are rarely consistent enough for a statistical method to be practically useful. In a context where one or more raters exhibit statistically inconsistent harshness (e.g., very large rater fit statistics) adjusting the scores of the examinees may be tricky. In the case of high stakes examinations, such statistical resolution techniques may prove to be difficult to support in courts, whereas, the blind scoring of a senior and experienced third rater may be easier to defend.

This study started with the use of a two-faceted Rating Scale Rasch model, and then a two-faceted Partial Credit model was applied to test the agency's assumption that the marking scheme was applied consistently across the four essays. Based on the results of the models this assumption did not hold. Then a three-faceted Partial Credit model was applied in order to accommodate for the differential severity of the raters as well as the differential difficulty of the essay questions. More complex many-faceted Rasch models could also be employed in order to accommodate for interaction effects between raters and questions. For example, by giving the model the freedom to model an idiosyncratic personal rating scale for each rater, it is very likely to lead to an even better fit of the data to the model.

However, the aim of this study was not to demonstrate all types of models that might be applied to datasets of this nature. The study aimed to illustrate how practitioners could use increasingly more sophisticated models by introducing additional meaningful facets and parameters in order to investigate how their questions function and how their mark schemes (rating scales) are applied. The noise in the data appeared as large fit statistics in the first two models, but this noise proved to be a result of the application of over-simplistic models; a vital facet (the raters) was actually missing from these models.

The question, however, is at what point should practitioners stop using increasingly more complex models to model their datasets? It is suggested that more sophisticated models are not needed if: (a) there is no strong theoretical evidence that the additional facets/parameters are meaningful, and (b) when there is a danger in turning the marking task into an infinite search for a theoretical ideal model-data fit. On the contrary, the aim should be to have a satisfactory fit for all practical intents and purposes, taking into account the context and the goals of the measurement exercise. It should be kept in mind that simpler models can sometimes be more defensible and easier to explain to politicians, parents, students, university recruitment agencies, and other stakeholders.

In the context of this study there were examinees who by sheer luck of the draw had their scripts marked by lenient raters and/or in some cases, made a lucky choice by deciding to respond to easier essay questions. Based on their scores, those examinees were granted access to further education. However, if the effects of the differential harshness of the raters and the differential difficulty of the

essays were taken into account when computing their scores, those students would not gain access to further education. By the same token, other students were not so lucky.

Less than optimal marking procedures, however, not only harm individuals but could also allow less able students to move to further education causing a suboptimal use of the limited public resources. It is frequently argued that in the cases of limited budgets examinations are held in order to filter out those examinees who are not predicted to perform satisfactorily in further education.

Why do a number of testing agencies and other exam-setting institutions around the world not use raw score adjustment methods? Although universities, books, journals, and the Internet make the knowledge easily accessible, policymakers and bureaucrats may not have the chance to learn and apply adjustment methods and to evaluate their merits. It could also be the case that policymakers and politicians lobby for the most intuitive and simple techniques to be used. The use of the raw scores is always intuitive, and the average everyday person will understand their usefulness and appreciate their simplicity. The use of complex statistical techniques that transform the raw scores so radically in an almost-magical way, may cause suspicion and doubt among some people. For example, it is very difficult to explain to parents or to examinees how a raw score that is almost twice as large as another score indicates approximately the same knowledge. Finally, every statistical adjustment requires the use of a relevant statistical model whose application demands that specific assumptions are met. For example, in order to retain the desirable psychometric properties of the Rasch model, it is necessary to verify that the data fit the model adequately for all practical intents and purposes of the assessment task. It may be the case, however, that not all the operational assessments fulfill all the assumptions for statistical score adjustment (e.g., for the application of a multifaceted Rasch model to the data).

However, it is true that merely adjusting scores for differences in the severity of the raters should not be considered a panacea. As it has been shown by relevant literature (Engelhard, 1994; Murphy, Jako, & Anhalt, 1993; Saal, Downey, & Lahey, 1980; Wolfe, Moulder, & Myford, 2001), a number of other rating issues such as the differential rater functioning over time (DRIFT) and idiosyncratic marking styles (e.g., halo effect, restriction of range) may affect marking. Adjusting for the differential severity of the raters will not deal with these issues. Moreover, the application of score adjustment techniques does not take into account the possibility that rater behavior may change over time. Recent research (Congdon & McQueen, 2000; Hoskens & Wilson, 2001; Wolfe et al., 2001) indicates that rater severity may not be stable across time and, therefore, more complex adjustments may be needed.

The methods employed in this study do not take a number of rating issues like the DRIFT and idiosyncratic marking styles (e.g., halo effect, restriction of range) into consideration. Before suggesting a blanket use of the methods presented in

this study, the effect of those issues on score adjustments must be investigated further.

Another limitation of the study is that the reasons that encouraged the examinees to choose one of the four essay topics over the other were not investigated. However, there is some meta-cognitive literature that suggests that smarter students choose easier prompts and those with less meta-cognitive ability will choose harder questions (Hopkins & Antes, 1990). To the extent to which this were true, then examinee ability would be confounded with prompt difficulty but the statistical methods employed in this study do not take this possibility into account. However, it is very difficult to see how this effect could have been accommodated statistically; the best way to differentiate between the difficulty of the essays and the ability of the students might be to set up an experimental design and ask the examinees to randomly answer one of the essays—therefore abandoning the self-selection of the essay questions.

CONCLUSION

This research used empirical data from a recent high-stakes national examination and investigated the impact of using increasingly more sophisticated models for the estimation of examinee ability. The models appeared to be efficient in adjusting the raw scores of the examinees to accommodate for the differential harshness of the raters and the differential difficulty of questions. Although the use of raw scores is intuitive and therefore easily understood by the stakeholders of education (e.g., parents, students, teachers, university recruitment boards), more desirable alternatives exist. It is up to policymakers to choose the appropriate model every time based on the context of each educational system.

One possible way to disseminate good practice would be to intensify the efforts to reach and explain the principles of fundamental measurement not only to academics and to practitioners but also, to some extent, to the wider public. In the case in which individuals are unfairly prevented from furthering their education because of the use of suboptimal marking methods, doing nothing does not seem to be an option.

REFERENCES

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.
- Banerji, M. (1999). Validation of scores/measures from a K-2 developmental assessment in mathematics. *Educational and Psychological Measurement*, 59, 694–715.
- Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, 2, 49–68.
- Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, 26, 364–375.

- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20, 89–110.
- Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, 13, 1–18.
- Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, 37, 163–178.
- Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5, 171–191.
- Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31, 93–112.
- Engelhard, G., & Myford, C. M. (2003). *Monitoring faculty consultant performance in the Advanced Placement English literature and composition program with a many-faceted Rasch model*. College Board Research Report: 2003—1. New York: College Entrance Examination Board.
- Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). London: McGraw-Hill.
- Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3, 49–68.
- Hopkins, C. D., & Antes, R. L. (1990). *Classroom measurement and evaluation* (3rd ed.). Itasca, IL: F. E. Peacock.
- Hoskens, M., & Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the Golden State Examination. *Journal of Educational Measurement*, 38, 121–145.
- Hurtz, G. M., & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement*, 59, 885–897.
- Johnson, R. L., Penny, J., & Gordon, B. (1999, April). *Score resolution and the inter-rater reliability of holistic scores*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
- Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and inter-rater reliability: An empirical study of an analytic rubric. *Applied Measurement in Education*, 13, 121–138.
- Johnson, R. L., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication*, 18, 229–249.
- Johnson, R., Penny, J., Fisher, S., & Kuhs, T. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education*, 16, 299–322.
- Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, 1, 152–176.
- Kingsbury, F. A. (1922). Analysing ratings and training raters. *Journal of Personnel Research*, 1, 377–388.
- Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 97–110). New York: Academic Press.
- Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.
- Linacre, J. M. (2006). *FACETS Rasch measurement computer program*. Chicago: MESA Press.
- Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement*, 3, 486–512.
- Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics*, 19, 171–200.
- Lumley, T., Lynch, B. K., & McNamara, T. F. (1994). A new approach to standard-setting in language assessment. *Melbourne Papers in Language Testing*, 3, 19–40.
- Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3, 331–345.

- Marcoulides, G. A. (1993). Maximizing power in generalizability studies under budget constraints. *Journal of Educational Statistics*, 18, 197–206.
- Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, 78, 218–225.
- Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement*, 3, 300–324.
- New South Wales Board of Studies (2002). *2002 HSC Update Newsletter 02*. Educational Resources, New South Wales, Australia. Accessed on November 30, 2006 at http://boardofstudies.nsw.edu.au/hsc/hsc_update_2002_02.html#adjusting
- Petersen, N. C., Kolen, M. J., & Hoover, H. D. (1993). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Phoenix, AZ: Oryx Press.
- Rae, G., & Hyland, P. (2001). Generalizability and classical test theory analyses of Koppitz's Scoring System for human figure drawings. *British Journal of Educational Psychology*, 71, 369–382.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413–428.
- Schlesinger, M. E. (1995). Is "No-Choice" the choice for future metallurgical PE Exams? *Journal of Outcome Measurement*, 47, 63–72.
- Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51, 541–565.
- Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, 1, 199–218.
- Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., DeKruif, R. E. L., Reed, M., Brown, T. T., Levine, M. D., & White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement*, 59, 492–506.
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25–29.
- Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263–287.
- Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6, 145–178.
- Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied Measurement*, 2, 256–280.
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
- Wright, B. D., & Mok, M. (2000). Rasch models overview. *Journal of Applied Measurement*, 1, 83–106.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

APPENDIX

In the context of this study, the goal of the equipercentile equating is to make the raw scores on two essays to correspond to the same cumulative percentage for a certain group of examinees (Angoff, 1984). The technique demands either the conversion of one set of scores to the other, or the conversion of the two sets of scores to a third (new) one.

For the equipercentile equating of two essays, one needs the graphs of the cumulative percentages for their scores. If the two graphs are drawn on the same axes, then the raw scores on each essay that correspond to the same cumulative percentage can be identified.

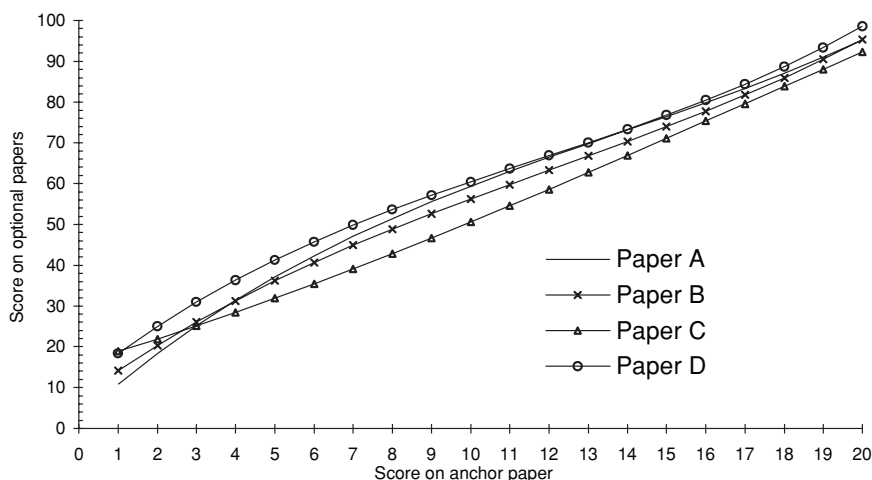


FIGURE 9
The equipercentile equating method for the four essays.

These scores form pairs of equivalent scores. Many such pairs can be plotted on a graph to form the conversion line, which is the function that transforms the scores on the one essay to the score scale of the other essay.

Although the testing agency uses the raw scores on the four essays as if they are comparable, the equipercentile method indicates that major differences between the essays exist. The equated scores for the four essays are shown in Figure 9.

Essay D is found to be the easiest because for the same score on the anchor essay the examinees receive a higher score on this essay rather than on any other essay. Essay A also appears to be fairly easy compared to the other two essays. Essay C is the most difficult essay of all. For example, a score of 55 on Essay C is equivalent to a score of approximately 64 on Essay D.

Copyright of International Journal of Testing is the property of Lawrence Erlbaum Associates and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.