Article

LANGUAGE TESTING

Application of an Automated Essay Scoring engine to English writing assessment using Many-Facet Rasch Measurement

Language Testing
2023, Vol. 40(1) 61–85

DOI: 10.1177/02655322221076025
journals.sagepub.com/home/ltj
SAGE

Kinnie Kin Yee Chan
Hong Kong Metropolitan University, Hong Kong

Trevor Bond
James Cook University, Australia

Zi Yan
The Education University of Hong Kong, Hong Kong

Abstract

We investigated the relationship between the scores assigned by an Automated Essay Scoring (AES) system, the Intelligent Essay Assessor (IEA), and grades allocated by trained, professional human raters to English essay writing by instigating two procedures novel to written-language assessment: the logistic transformation of AES raw scores into hierarchically ordered grades, and the co-calibration of all essay scoring data in a single Rasch measurement framework. A total of 3453 essays were written by 589 US students (in Grades 4, 6, 8, 10, and 12), in response to 18 National Assessment of Educational Progress (NAEP) writing prompts at three grade levels (4, 8, & 12). We randomly assigned one of two versions of the assessment, A or B, to each student. Each version comprised a narrative (N), an informative (I), and a persuasive (P) prompt. Nineteen experienced assessors graded the essays holistically using NAEP scoring guidelines, using a rotating plan in which each essay was rated by four raters. Each essay was additionally scored using the IEA. We estimated the effects of rater, prompt, student, and rubric by using a Many-Facet Rasch Measurement (MFRM) model. Last, within a single Rasch measurement scale, we co-calibrated the students' grades from human raters and their grades from the IEA to compare them. The AES machine maintained equivalence with human scored ratings and were more consistent than those from human raters.

Keywords

Automated Essay Scoring (AES) system, English essay assessment, FACETS, human raters, Intelligent Essay Assessor (IEA), Many-Facet Rasch Measurement (MFRM)

Corresponding author:
Kinnie Kin Yee Chan, Hong Kong Metropolitan University, 30 Good Shepherd Street, Homantin, Hong Kong.
Email: kiyechan@hkmu.edu.hk

多面Rasch測量(MFRM)模型應用於自動文章評分(AES)系統作英文作文評估

Translated Abstract
我們通過研究自動文章評分系統(AES)的文章評分器(IEA)分配的分數和由訓練有素的專業人工評分員分配的英語寫作分數之間的關係，我們提出了兩個新的寫作評估程序: 將AES原始分數轉換成次弟分數，以及在單一Rasch測量框架中對所有論文分數進行協同校準。數據來自於589名美國學生 (4、6、8、10和12年級)所寫的3453篇文章。所有文章來自於國家教育進展評估(NAEP)其中三個年級(4、8和12年級)的寫作題目。我們隨機分配給每個學生兩個版本的評估，A或B。19位經驗豐富的評審員使用NAEP評分準則對文章進行整體評分，每篇文章由IEA以及4位評分員評分。我們使用多面Rasch測量(MFRM)模型估計了評分者、題目、學生對評分的影響。最後，在一個單一的Rasch測量量表中，我們共同校準了人工評分和IEA的評分，以進行比較。AES機器保持與人類評分基本相若，並且比人類評分更一致。

Translated keywords
多面Rasch測量模型(MFRM) 、自動文章評分系統(AES) 、文章評分器(IEA) 、人工評分員、英文作文評估、FACETS

Introduction

In both classroom and high-stakes testing, essay writing is considered to be a form of performance assessment (Bachman, 1990). If reliable and accurate assessments can be obtained that are consistent in scoring student performances, teachers' and other stakeholders' confidence in those results will be enhanced and those assessment data can be aggregated or combined effectively across classrooms, grade levels, departments, schools, or districts (Gottlieb, 2006).

Judging a student's writing can be a complicated task (Hamp-Lyons, 2019; Linacre, 1989). How is any essay given the mark that it is worth? How do all essays of an equivalent worth get the same mark? These are the questions still routinely being asked, despite the already substantial discussion around fairness and consistency of essay scoring in educational assessment (Banerjee, 2017; Chowdhury, 2020; Wheadon et al., 2020).

In reality, marking discrepancies are almost inevitable. In the assessment of writing, differences between scores assigned by different essay raters are still a serious concern, even with experienced and proficient raters considering all criteria equally to justify their ratings (Lim, 2011). Studies of rater scoring indicate that experienced raters not only implement strategies based on the assessment criteria for an essay, but also on their own behaviors during actual rating sessions (Cumming, 1990). Even when raters judge the same essay against the same scoring rubrics, they do not usually have perfect agreement on scores (Erguyan & Aksu Dunya, 2020; Trace et al., 2017). The backgrounds and qualifications of the raters might affect their understanding of the marking rubrics (Eckes,

2008), the different genres of the prompts (Weigle, 1999), and the suggested criteria for a piece of good writing (Schaefer, 2008).

Fortunately, the implementation of Many-Facet Rasch Measurement (MFRM) (McNamara, 1996, 2011) has revolutionized essay examination marking and research by allowing examination authorities to adjust for variations in essay difficulty, and to identify and manage variations in rater characteristics, including relative severity or leniency, lack of consistency in scoring, the influence of rater training and professional background, as well as consistency in the scoring of essay writing over time (Engelhard, 1994; Linacre, 1989; McNamara & Knoch, 2012). It does so by building a single and complete frame of measurement reference in which all influential elements are estimated simultaneously in order that each can be inspected individually (Linacre, 1989).

Automated Essay Scoring (AES) systems have been proposed as supplements to human rating assessment (Rupp et al., 2019; Shermis & Burstein, 2013). In the 1960s, the first automated writing evaluation system was developed by Ellis Page (1966). As substitutes for human raters, AES systems utilize computer technology to evaluate and score written prose. At the very minimum, AES systems provide consistency in scoring essays, and are not time-consuming for scoring routine, written-language assignments (Ifenthaler & Dikli, 2015; Rotou & Rupp, 2020).

Numerous studies across decades have investigated the effectiveness and appropriateness of a multitude of AES systems for writing assessment (Xi, 2010). Moe (1980) claimed that text analysis by computers had been feasibly and practically demonstrated. Shermis and Burstein (2003) reported positive correlations between AES scores and human raters, thereby claiming a high level of construct validity. Rudner and Gagne (2021) asserted in 2001 that substantial success in using such software to score essays had already been achieved. High reliability of three software systems, Project Essay Grade (PEG) (Page, 2003), Intelligent Essay Assessor (IEA) (Pearson, 1998), and the e-rater (Burstein & Kaplan, 1996), was demonstrated in that study. Consequentially, they advised schools and state educational systems in the United States to consider using automated scoring services in writing assessments. Barrett's (2015) comparison of ratings of Criterion and human raters on a writing assessment led to his conclusion that computer ratings correlated well with human ratings ($r = .64$). Hoang (2011) correlated the scores of Intellimetric and human raters ($r = .69$). In a set of 188 essays in the Graduate Management Achievement Test (The Graduate Management Admission Council, n.d.), or GMAT, IEA scores and scores assigned by human raters correlated well ($r = .80$) (Foltz et al., 1999). It seems reasonable to conclude that computers could be useful for furthering the versatility and efficiency of written essay assessments scoring.

Although the scores generated by AESs are routinely close to the scores of human raters in writing assessments (Correnti et al., 2020; Rudner et al., 2006; Warschauer & Grimes, 2008), AESs are often touted as instruments to supplement, rather than to replace expert human raters. For instance, an AES engine is used by the GMAT Analytic Writing Assessment in parallel with one human rater to generate two scores for each essay. Any essay is then reviewed only if there is a disparity between the machine score and the human rater score, in order to settle the difference and confirm the final score (GMAT). Nevertheless, by offering a solution to help avoid possible biases in low-stakes or high-stakes assessments, English language teachers might see AESs as a practical and

trustworthy tool to lower the assessment workload, and to release some of the pressure from grading student essay writing.

AES methods for English language written essay assessments have become commonly recognized in both low-stakes classroom assessments and large-scale high-stakes standardized tests in the United States (Dikli, 2006; Williamson et al., 2012). Yet, little empirical evidence beyond correlation analyses is in place to corroborate the performance of the AES models, and only some work has been completed using Rasch measurement to calibrate human rater scores against AES scoring. We aimed to examine the relationship between one AES system, the IEA, and human scoring of English essay writing. Rather than adopting the comparison methods previously used in this type of research (Williamson et al., 2012), we implemented two innovative practices: First, we transformed the IEA raw scores into nine hierarchically ordered grades, and second, we developed a single, comparative measurement scale, using the Rasch model, to co-calibrate the IEA transformed grades with the students' human-rated MFRM measures. The use of MFRM in this research is unique because we include computer-generated raw scores in the analytical model.

AES system in this research

IEA

Pearson Education purchased the IEA, which was initially developed by the University of Colorado, in the late 1990s (Warschauer & Ware, 2006). Based on the Pearson's Knowledge Analysis Technologies (KAT) engine, the internet-based AES system, IEA, scores the content of essays. The Latent Semantic Analysis (LSA) incorporated into the IEA is a statistical language learning concept that obtains information about a specific topic with 50,000 to 10 million words from sources on the internet, and estimates the semantic similarity of words and essays by evaluating a large corpus of relevant text (Pearson, 2019a).

Landauer et al. (2003) claimed that the LSA scores student essays so as to identify the meaning of the contained words and then compares that with essays of known quality on relevant ideas and concepts. By comparing an essay with a set of essays previously scored by human raters, the IEA engine then evaluates the content of any similar essay. The IEA engine needs between 100 and 300 previously rated sample essays for training and calibration purposes, for any particular topic or prompt. Those essays must have been rated by two independent human raters, and then by a third human rater if the first two did not reach consensus. The scores from these human raters are then used to train the IEA. The IEA score assigned to each essay is based on the essay's similarity in the content to those from the training sets (Pearson, 2019a). Wohlpart et al. (2008) found that the software operates most effectively with very narrowly prescribed prompts on essays between 100 and 500 words in length. A mathematical representation of the relations among words and passages is created by the system of statistical computations, which can then be employed to assess a large number of similar essays based on the particular prompt. Furthermore, Foltz et al. (2014) reported that the IEA scored essays more consistently than did human raters, while having 81% of scores matching those of human raters. The IEA can provide assessments of the words in any subject area and can

incorporate built-in detectors to alert human raters in case there are off-topic responses, or other circumstances requiring human raters' intervention (Pearson, 2019b).

For academic papers in the field of education, if essays are very similar in their use of words and the writing styles, plagiarism can also be inferred by the IEA, especially when grading large numbers of essays in an assessment (Landauer et al., 2003). But the interpretation of syntax, grammar, style, or mechanics cannot be analyzed by the LSA system; nor can some specific genres, such as rhymes and poems.

MFRM. The Many-facet Rasch model (Linacre, 1989), an extended version of the original dichotomous Rasch model for measurement, is routinely used to analyze the judgments of human raters in writing assessments to provide fairer performance scores. McNamara (1996) claimed that MFRM provides a well-founded technique to help ensure the validity of the assessment by measuring various related influential facets including the interactions among students, raters, and prompts. Furthermore, the MFRM can reveal the extent of disagreement among raters, and can provide empirical information to assist raters to achieve self-consistency.

The following Many-facets Rasch model is typical for a writing assessment analysis with three facets, namely, the student's writing ability (B), the prompt difficulty (D), the rater's severity (C), and the rating scale step difficulty (F)

$$\log\left(P_{nijk} / P_{nij(k-1)}\right) = \mathrm{B}_n - \mathrm{D}_i - \mathrm{C}_j - \mathrm{F}_k$$

where P_{nijk} is the probability of student n being awarded on prompt i by rater j a rating of k; $P_{nij(k\text{-}1)}$ is the probability of student n being awarded on prompt i by rater j a rating of $k - 1$; Bn is the ability of student n; D_i is the difficulty of prompt i; C_j is the severity of rater j; and F_k is the difficulty of the step from category $k - 1$ to category k and $k = 1$ (Bond et al., 2020, p. 314).

MFRM can clearly quantify the differences in rater severity of essay scoring: the consistency of raters can be estimated in an MFRM analysis, and the results can be used to adjust students' essay scores for the differences in prompt difficulty and rater severity in the scoring process.

In this research, we applied a double-marking rotation plan to have four out of the 19 trained and paid raters grade each essay in the data set in order to monitor the rater severity effect. The rater pairings were systematically changed, and the repeated design was used to generate the rating plans. This MFRM design is tailored to monitor the rater effect in the assignment of double-marking, which is consonant with the calculation of rater effect estimations (Bond et al., 2020).

Research question

The primary research question guiding the current study is:

What is the relationship between computer scoring and human scoring of English essay writing?

To answer this question, we applied the Rasch model to develop a comparative measurement scale for calibrating the IEA scores against the students' MFRM grades as judged by the human raters.

Method

Participants

The data for this study came from the dissertation of the first author (Chan, 2012); in that dissertation, two different sets of data were used, from Hong Kong and the United States, respectively. The research involving the data set of Hong Kong students was reported in a book chapter (Chan & Bond, 2016). This current research is based on the data set from students in the United States. The first author collected 3453 essays from 589 students (Grades 4, 6, 8, 10, and 12) who were L1 speakers of English in one school district in north-central Mississippi of the United States. The gender information of students, provided by the school district, comprised 307 females (52.1%) and 282 males (47.9%). All of the students were administered four to six writing prompts. Then, every essay was rated by four of the 19 independent trained raters, and scored with the IEA.

Instruments—prompts for essay writing

The 18 selected prompts used in this research were posted for public use on the website of the National Assessment of Educational Progress (NAEP) (US Department of Education, Institute of Education Sciences [USDOE], 1998, 2002). Over a 3-week period, students in each grade (Grades 4, 6, 8, 10, and 12) wrote two essay responses per week, covering a total of six prompts across narrative, informative, and persuasive genres (USDOE, 1998).

Table 1 presents the prompt design by grade. In order to provide linkage across the grades, two extra grades of students, 6 and 10, were included in the prompt design. Three prompts from the grade above and another three prompts from the grade below were taken by each Grade 6 and Grade 10 student. A linked series for the essay writing scoring system was provided by the linkage prompts.

As explained previously, the two versions, A and B, were randomly assigned to students. Each comprised a narrative (N), an informative (I), and a persuasive (P) prompt. A total of 30 minutes (5 minutes for planning and 25 minutes for writing) were allowed for students to write to each prompt. Different prompts were distributed to students and each class would answer to non-identical prompts on any given day, in order to nullify ordering and learning effects.

Model

In the current research, five defined facets for the analysis are notated in the formula below as follows: the writing ability of the student, the difficulty of the prompt, the severity of the rater, the severity of the AES system, and the difficulty of the rating scale step. Because a holistic

Table 1. Prompt design by grade.

Grade	Prompts					
4	4AN	4AI	4AP	4BN	4BI	4BP
6	4BN	4BI	4BP	8AN	8AI	8AP
8	8AN	8AI	8AP	8BN	8BI	8BP
10	8BN	8BI	8BP	12AN	12AI	12AP
12	12AN	12AI	12AP	12BN	12BI	12BP

Note: A and B are two different prompt indicators. N, I, and P stand for narrative, informative, and persuasive, respectively.

(rather than analytic) rubric was used to score the essays, the domain facet is not included in this model (Linacre, 1989). Thus, the formula for the partial credit MFRM model is

$$\log\left(P_{nijkx} / P_{nijk(x-1)}\right) = B_n - D_i - C_j - A_k - F_x$$

where P_{nijkx} is the probability of student n being rated, on prompt i by rater j and AES k, a rating of x; $P_{nijk(x-1)}$ is the probability of student n being rated, on prompt i by rater j and AES k, a rating of x-1; Bn is the writing ability of student n; D_i is the difficulty of prompt i; C_j is the severity of rater j from the United States; A_k is the severity of IEA; and F_x is the difficulty of the rating step up from category x-1 to category x.

In this Many-facets Rasch model, facets of the student essays can be modeled and their effects on the scoring of the raters and the AES system are estimated. There is no pre-set limit to the number of facets that can interact to produce a rating, with the caveat that each facet should provide a particular meaning in the research (Linacre, 1989). The elements of each facet are summarized by the measure mean, the standard deviation, the reliability of element separation, and the corresponding chi-square for homogeneity (Lunz et al., 1990).

Scoring procedures

An established testing company employed raters with extensive essay scoring experience. Raters assigned essay scores by using published holistic scoring guidelines, which were set in line with the National Assessment of Educational Progress (NAEP) testing design. NAEP is the assessment implemented to assess US students' knowledge in various school subjects including mathematics, reading, science, and writing. Consequently, schools might design school-based curricula for their students based on NAEP results. In this research, unique scoring guidelines for each prompt in a particular genre were adopted for scoring the prompts at each grade level. The same 6-point rating scale was used to report the grading of all essays in the genres of narrative, informative or persuasive, but each prompt had an unique scoring system, in that the level descriptors for each level were different for each different prompt. Overall, a scale with six ratings was used to grade the essays: 1—unsatisfactory, 2—insufficient, 3—uneven, 4—sufficient, 5—skillful, and 6—excellent.

Results

Human raters

The data matrix for the essays scored by the human raters contained 24 ratings of 1–6 for each student (four ratings from four raters for each of six essays) with grade level, prompt, and genre recorded for each.

Transformation of raw scores from the IEA

The raw scores generated by the IEA engine ranged from 1 to 100. Although a large number of response categories (up to 255 categories) can be accepted for Rasch analyses using WINSTEPS, a much more reasonable number of categories would assist with the meaningful interpretation of the results, and allow departures from fit to the model to be detected more clearly (Linacre, 1989). Thus, a Poisson logarithmic transformation (Bond et al., 2020; Yan & Bond, 2011) was applied to the raw score data from the AES system with the aim of creating a reasonable range of meaningful ordered response categories.

The transformation can be expressed as

$$\text{Scored category} = 1 + n^{*}\frac{\log(\text{observation}+1) - \log(L+1)}{\log(H+1) - \log(L+1)} \quad (1)$$

where L is the lowest, and H is the highest observed raw score. For example, when the intended transformation target was a nine-category scoring structure, 8 was chosen as the n value, with nine eventually being regarded as a reasonable number of hierarchically ordered scoring categories.

This Poisson logarithmic transformation broke the original raw score observations into a certain number of intervals, with all observations within any one interval classified into the same category level. Hence, at the first stage of the analyses in this research, the IEA raw scores of the student essays were iteratively reduced from 20 to 9 ordered categories (Figures 1 and 2) in order to produce more even distributions of grades and more understandable category structures.

The first iteration showed the IEA results for the 20 categories transformation had 93% of all data in Categories 16–19 (i.e., only four categories). Figure 1 shows that Categories 1–15 and Category 20 to be entirely collapsed and without specific peaks, meaning that those categories were actually redundant. Categories with low frequencies are problematic for analysis because there are not enough observations to allow for an estimation of stable category threshold values (Bond et al., 2020; Linacre, 2002). The results for the initial nine category transformation (Figure 2) are similarly inadequate.

Linacre (1999, 2002) stated that a regular uniform distribution across categories should provide more satisfactory Rasch estimates. The trial transformations of the IEA raw scores were used for the student essays. Therefore, the category structure of the AES system needed further investigation and reconstruction.

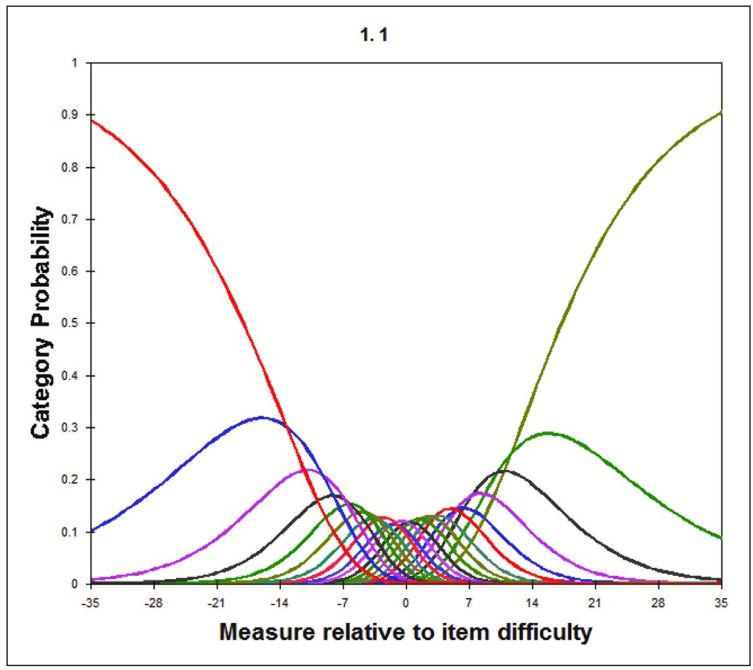

Figure 1. Category probability curves for the IEA scores on the 4AI prompt (20 categories).

Collapsing of categories

As an ideal categorization of the raw data could not be found by using the Poisson logarithmic transformation alone, the next step followed the Bond et al. (2020) suggestion that collapsing categories is a productive way to investigate the best match between Rasch analysis diagnostics and respondent use. This was done to help develop the final phase of scoring with the most appropriate categorization.

The Linacre guidelines for combining categories were applied to produce a measurement process under the condition of an equal contribution for each category in the Rasch analysis (Linacre, 1995, 1999, 2002).

Figure 3 shows the results of the re-categorization of the nine new category-IEA data. The category probability curves for the IEA scores of the student essays reflect that the categories are evenly distributed with specific peaks for each of the nine new categories.

The MFRM analyses results

In order to answer the primary research question, it is necessary to undertake a successful MFRM analysis of the human scoring of those student essays; then to construct a comparative measurement scale for calibrating the nine-category IEA grades against the students' MFRM judged ratings.

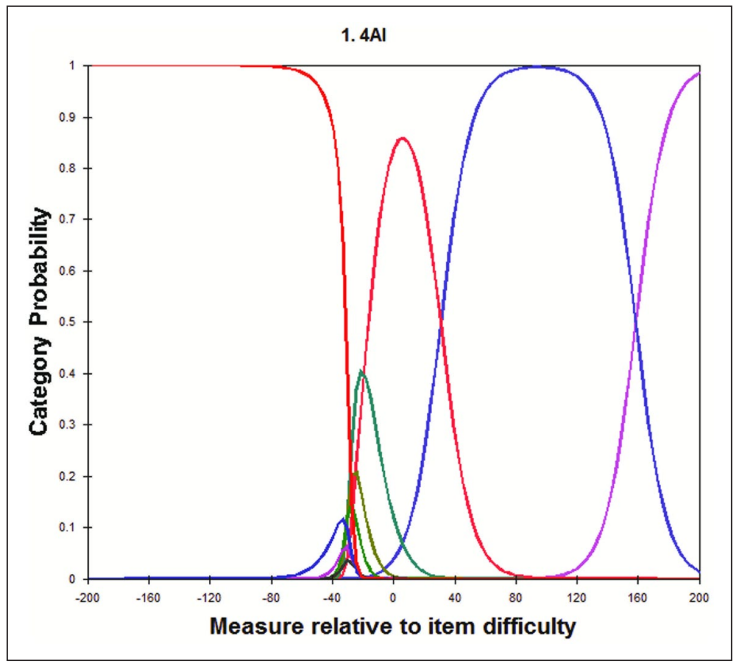

Figure 2. Category probability curves for the IEA scores of the 4AI prompt (9 categories).

Investigating the impact of various facets in the MFRM analysis of these data should lead to a deeper understanding of each of the aspects involved in the English essay assessment setting (McNamara, 2011). Generally, the results of the MFRM analysis of the human scoring of the student essays in this research should be informative to those ends. The performances of the eighteen writing prompts and the nineteen human raters fit the Rasch model's measurement requirements very well, notwithstanding that the overall performances of 33 out of 589 students showed underfit to the Rasch model with infit Zstd > +2 (Linacre, 1989).

Figure 4 maps the Rasch measurement of the difficulty of 18 prompts, the severity of 19 raters, and the ability of 589 students along with the IEA response category thresholds. In this MFRM variable map, column 1 shows the logit interval measurement scale, which extends from −7 to +5 logits, and column 6 presents the NAEP holistic rating criteria, which range from 1 to 6. In columns 2 through 5, 0.00 logits was adopted, by default, as the mean estimate of prompt difficulty, rater severity, and student ability.

The most difficult prompt, 12AP (Grade 12 NAEP Persuasive prompt, Version A), was located on the top of the scale (+2.25 logits). Grades 10 and 12 students were assigned to write 12AP in this research. The easiest prompt was persuasive 4AP (−2.90 logits) and was used for Grade 4 students only. In column 3 (Raters) of the scale, higher measures indicate raters who were more severe than other raters in essay scoring: R200 was the most severe rater (+1.28 logits), and R108 the most lenient (−1.75 logits). In column 4 (IEA), the thresholds for the machine essay scoring categories were ordered

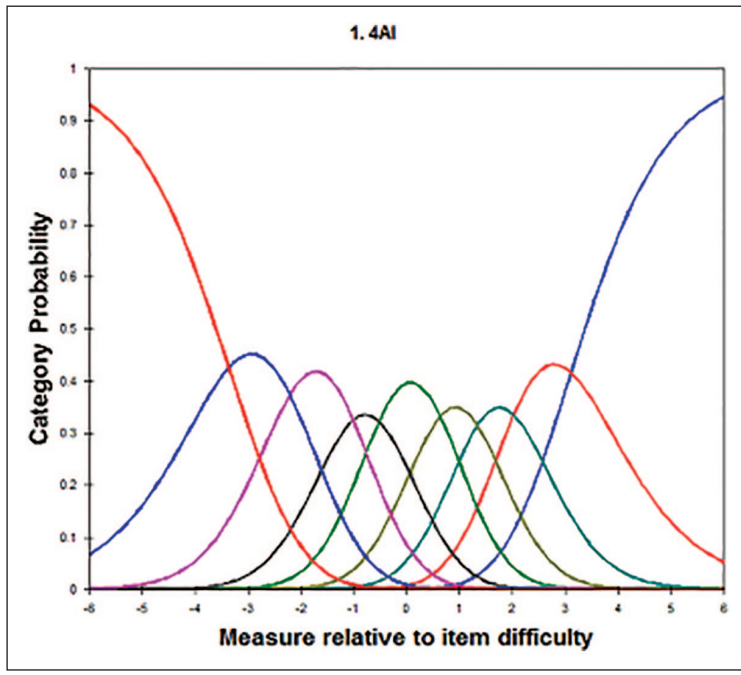

Figure 3. Category probability curves for the IEA scores on the 4AI prompt after collapsing (new nine categories).

appropriately, 1 through 9, with the range for the IEA from −4 to +4 logits. The overall essay writing ability measures of 589 students (column 5) ranged from a low near −6 to the most capable students located at nearly +4 logits, with most of the students' writing abilities (well over 90%) estimated between −2 and +2 logits.

Calibration of students. Table 2 shows the Rasch person statistics for the 33 under-fitting students (5.6% of total number of 589 students). The under-fitting students might have performed erratically or produced off-topic essays when they responded to particular questions. The essays of nine under-fitting students were scored by the most severe or lenient rater(s) in this research.

The lowest writing ability (Grade 4 student) is located at −6.33 logits and the highest writing ability (Grade 12 student) is at +4.17 logits, revealing a wide range of essay writing ability with more than 10 logits between Grades 4 and 12 students.

Moreover, a number of under-fitting students were identified. As under-fitting persons (Zstd > +2.0) indicate more problems with the quality of the measurement scale than over-fitting persons do, all under-fitting students were identified as having Zstd greater than +2. These included 15 Grade 4 students, 9 Grade 6 students, 4 Grade 8 students, 3 Grade 10 students, and 2 Grade 12 students.

The results for two under-fitting exemplar students (579 and 366) show (qualitatively) how they performed unexpectedly when responding to prompts, thereby producing

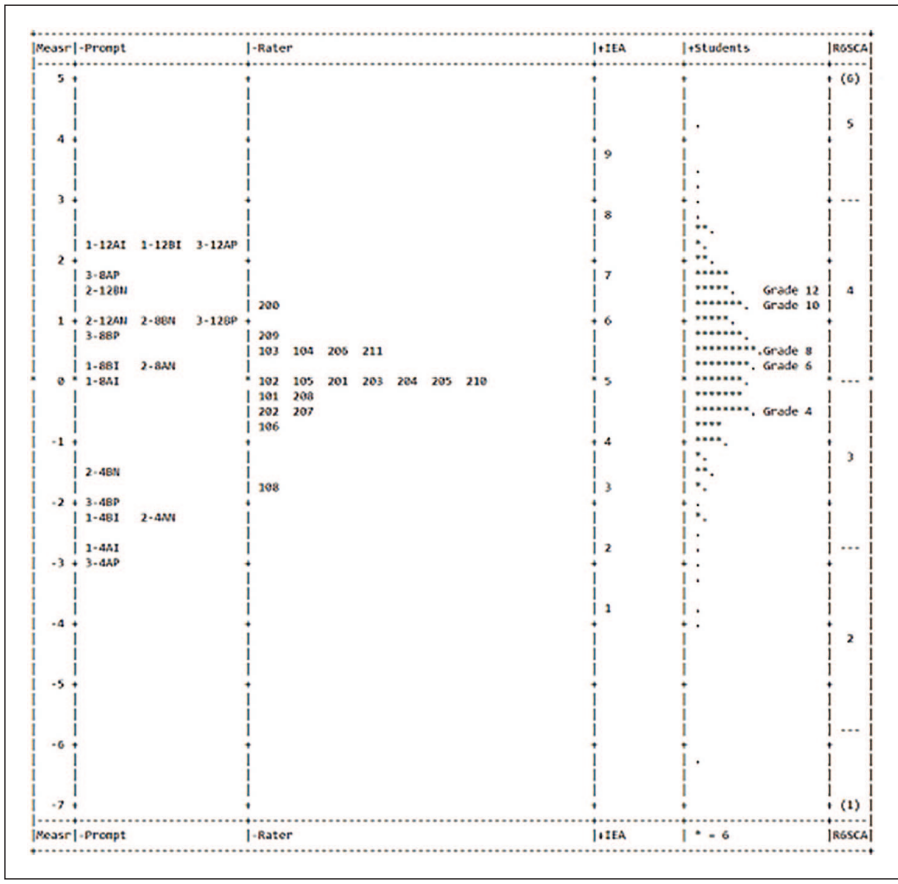

Figure 4. Rasch variable map for essays scored by human raters and IEA.
*In the column "Students," Grades 4–12 annotations indicate the grade means of Rasch measures.

quantitative indicators (Rasch fit statistics), which flagged their essays and grades as anomalous, or requiring closer inspection by human raters.

Student 579 (infit mean square 8.44, Zstd 9.0; outfit mean square 7.62, Zstd 9.0) whose performances exhibited a noisy (erratic) scoring pattern had written six essays in total. Table 3 shows that the rating pattern of six different combinations of four raters for this student's essays.

This student, Student 579, seems to have misinterpreted the instructions in prompt 4AP; this student described "a happy life with families" rather than writing "a letter to convince a friend to be visible," which was the requirement of the 4AP prompt. Consequently, Student 579 obtained erratic scores on that prompt: the ratings for that prompt were much worse than his average performance across the other essay prompts.

One more student, Student 366 (infit mean square 2.83; Zstd 4.1; outfit mean square 2.79; Zstd 4.0), wrote six essays in total and produced an off-topic essay when he

Table 2. Rasch measures and Zstd of 33 under-fitting students.

Student		Observed score	Measure (Logit)	*SE*	Infit		Outfit	
Number	Group				MnSq	ZStd	MnSq	ZStd
135	108	5.29	3.44	0.34	1.64	2.1	1.71	2.3
109	106	4.54	1.59	0.33	1.67	2.1	1.64	2.0
292	106	5.21	2.46	0.36	1.68	2.1	1.39	1.0
513	112	3.71	0.61	0.34	1.71	2.1	1.71	2.1
288	104	4.50	0.84	0.33	1.76	2.2	1.79	2.3
539	112	4.88	2.34	0.33	1.74	2.3	1.71	2.3
357	110	4.08	1.18	0.34	1.83	2.4	1.87	2.5
237	106	3.00	–1.51	0.35	1.85	2.4	1.86	2.4
269	104	4.96	1.44	0.32	1.80	2.5	1.80	2.6
031	106	3.50	–0.88	0.34	1.90	2.6	1.88	2.5
195	108	3.60	–0.41	0.38	2.01	2.6	2.01	2.6
224	104	2.44	–1.64	0.42	2.11	2.6	2.11	2.6
188	108	3.29	0.98	0.35	1.94	2.7	1.94	2.7
052	104	3.17	–0.99	0.35	1.99	2.8	2.00	2.8
248	106	3.79	0.76	0.34	2.00	2.8	1.96	2.7
058	104	2.46	–2.16	0.34	2.00	2.8	1.99	2.8
451	106	4.00	0.37	0.34	2.04	2.8	2.07	2.8
306	106	3.00	–1.81	0.42	2.31	2.8	2.38	2.8
101	108	2.92	–0.66	0.34	2.04	2.9	2.04	2.9
021	104	3.33	–1.65	0.35	2.23	3.2	2.24	3.2
362	110	3.29	–0.76	0.35	2.25	3.3	2.25	3.2
581	104	3.13	–1.63	0.35	2.35	3.4	2.36	3.4
017	104	1.55	–3.84	0.41	2.31	3.6	2.29	3.6
281	104	4.17	–0.71	0.34	2.52	3.7	2.62	3.8
299	106	3.80	0.46	0.37	2.78	3.9	2.95	4.1
214	104	2.88	–2.36	0.34	2.61	4.1	2.61	4.1
366	110	3.88	1.24	0.35	2.83	4.1	2.79	4.0
465	106	2.33	–3.22	0.36	2.81	4.3	2.70	3.9
572	104	3.17	–1.20	0.35	2.85	4.4	2.84	4.4
049	104	2.63	–2.08	0.34	3.17	4.9	3.19	4.9
577	104	3.00	–2.09	0.35	4.23	6.5	4.22	6.5
271	104	3.15	–1.46	0.38	4.88	6.6	4.90	6.6
579	104	3.83	–1.28	0.34	8.44	9.0	7.62	9.0
Separation: 3.43			Reliability: 0.92					
Fixed (all same) chi-square: 7214.5 significance *p*: .00								

Note: SE: standard error; 104: Grade 4 student; 106: Grade 6 student; 108: Grade 8 student; 110: Grade 10 student; 112: Grade 12 student.

responded to the 12AN prompt on Day 6. Table 3 shows that the rating pattern of six different combinations of 4 raters on Student 366's essays. Students were asked to write a story about a special object in the 12AN prompt. They were required to describe the

Table 3. The rating pattern of six different combinations of four raters, the measures of the IEA for two under-fitting students' essays.

	Student 579						Student 366					
Day	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Prompt	4AP	4AI	4AN	4BN	4BI	4BP	8BI	8BP	8BN	12AI	12AP	12AN
Human ratings	1,1,1,1	4,4,3,4	5,4,4,4	5,5,4,5	4,4,4,4	6,5,5,5	5,6,6,5	5,5,4,3	5,5,4,5	4,4,4,4	4,4,4,2	2,1,1,1
IEA measure	62	43	50	57	48	52	64	59	66	68	67	33

Note: IEA: Intelligent Essay Assessor.

first encounter of the main character with the object, and to explain how important of the object to the character. An excerpt from Student 366 on the 12AN prompt follows:

> Once upon a time I felt very sad during the particular day on which we do these papers. Then I thought to myself, "Wow, I've done pretty well so far, but today's paper is just not gonna happen. I only need enough writing here to fill up this and part of the second to seem like I'm accomplished for the day." So in conclusion, if you've read this far, I'm just gonna do it right now.

For the performance of Student 366 on the 12AN prompt, no matter the human ratings and machine measures, both were much worse than his above-average performances on the other five prompts. This student seems to have produced an off-topic essay on purpose in this piece of writing.

Calibration of prompts. The human ratings (only) of the essays with the MFRM analyses revealed that all human-rated prompts fit the Rasch model very well, but prompts 4AP and 4BP, both persuasive prompts, showed misfit in the results of the IEA scoring. This is unusual and remains to be investigated. We provide more information in the next paragraph.

In Table 4, the calibrations for the 18 prompts in three genres are presented. The range of prompt difficulties is 5.15 logits, from the easiest prompt, 4AP (−2.90 logits), to the most difficult prompt, 12AP (+2.25 logits). Only two moderately misfitting prompts were found in this analysis; both were Grade 4 Persuasive prompts: 4AP (infit mean square 1.79; outfit mean square 1.75), and 4BP (infit and outfit mean square 1.54). In this research, 4AP was used for Grade 4 students and 4BP for Grades 4 and 6 students; for 4AP, students were requested to write a persuasive letter (4AP) to convince a friend to become visible, and to write another persuasive letter with details, examples, or reasons (4BP) to the school librarian to make the librarian buy a missing favorite book. Perhaps the human raters made generous professional adjustments in their scoring of these two prompts, given that they might think that these persuasive prompts were a challenge for Grade 4 students, especially given that these young students were required to build some strong possible arguments to persuade the target reader with the identification of the most convincing evidence to accept a particular point of view or take a specific action.

Table 4. Calibration of the prompts for student essays using human raters and IEA.

Prompt	Observed score	Measure (Logit)	*SE*	Infit		Outfit	
				MnSq	ZStd	MnSq	ZStd
12BI	3.45	2.15	0.09	0.92	–1.0	0.92	–1.1
12BN	4.04	1.39	0.08	0.85	–2.1	0.84	–2.3
12BP	4.02	0.99	0.09	1.12	1.5	1.11	1.4
12AI	3.77	2.20	0.06	0.86	–2.8	0.85	–2.9
12AN	3.94	0.96	0.06	0.77	–4.7	0.78	–4.6
12AP	3.37	2.25	0.06	0.91	–1.8	0.91	–1.8
8BI	3.77	0.18	0.06	0.98	–0.4	0.97	–0.6
8BN	3.70	1.02	0.06	0.71	–6.7	0.71	–6.6
8BP	3.64	0.84	0.06	1.23	4.4	1.22	4.3
8AI	3.48	0.00	0.05	1.15	3.1	1.15	3.1
8AN	3.71	0.20	0.05	0.71	–7.2	0.70	–7.4
8AP	2.90	1.75	0.05	0.95	–1.2	0.95	–1.2
4BI	3.74	–2.34	0.05	1.10	2.2	1.10	2.1
4BN	3.76	–1.58	0.05	0.74	–6.5	0.75	–6.1
4BP	3.60	–2.02	0.05	1.54	9.0	1.54	9.0
4AI	3.59	–2.85	0.07	0.87	–2.1	0.87	–2.2
4AN	3.69	–2.24	0.07	0.84	–2.6	0.90	–1.6
4AP	3.61	–2.90	0.07	1.79	9.0	1.75	9.0
Separation: 27.40			Reliability: 1.00				
Fixed (all same) chi-square: 14034.1			significance *p*: .00				

Human raters and AES essay ratings. The MFRM analyses show that all 19 human raters have acceptable fit to the model (infit and outfit mean square between .61 and 1.38) indicating that each of the professional essay raters scored the essays consistently and predictably. All the fit statistics for the IEA scaled essay scoring categories 1–9 are very close to the modeled value of 1.0, providing strong evidence that the IEA engine scored student essays very consistently.

In Table 5, the severity span across the raters (R200 to R108) is 3.03 logits. The relatively large standard errors calculated for raters R108 (0.39) and R211 (0.40) may be due to the fact that they scored only 19 and 18 essays respectively, out of the 3453 essays in the data set.

Calibration of IEA measures against human ratings. Table 6 presents the calibrations for IEA scoring. Category 1 is estimated at −3.82 logits, and Category 9 is at 3.87, showing a wide range of 7.69 logits between the lowest and highest categories. For IEA, the separation index is 49.88, which corresponds to a high separation reliability of $R = 1.00$. The fixed chi-square is 17878.6 ($p < .001$), showing categories have different measures. A high reliability of separation index of 1.00 indicates that some categories are statistically significantly higher, and others are statistically significantly lower, in the analysis. The standard errors are between .03 and .07 and show a high precision of measurement.

Table 5. Calibration of the raters for student essays using human raters and IEA.

Rater	[a]Essays scored	Observed score	Measure (Logit)	*SE*	Infit		Outfit	
					MnSq	ZStd	MnSq	ZStd
R101	894	3.86	–0.28	0.06	0.77	–5.1	0.80	–4.4
R102	1002	3.79	–0.05	0.05	0.74	–6.3	0.74	–6.2
R103	297	3.60	0.49	0.10	0.71	–3.8	0.70	–3.8
R104	976	3.58	0.55	0.05	0.70	–7.2	0.71	–6.8
R105	922	3.79	–0.05	0.06	0.91	–1.8	0.91	–1.9
R106	621	4.04	–0.76	0.07	0.63	–7.5	0.63	–7.4
R108	19	3.32	–1.75	0.39	0.62	–1.2	0.61	–1.2
R200	607	3.18	1.28	0.07	1.28	4.5	1.28	4.4
R201	998	3.60	–0.07	0.05	1.23	4.7	1.22	4.4
R202	626	3.60	–0.41	0.07	1.20	3.3	1.19	3.1
R203	1133	3.56	0.07	0.05	1.20	4.4	1.20	4.4
R204	1013	3.65	0.08	0.05	0.95	–1.1	0.95	–1.1
R205	768	3.49	0.04	0.06	1.08	1.4	1.08	1.4
R206	988	3.42	0.41	0.05	0.87	–2.9	0.87	–2.9
R207	777	3.66	–0.42	0.06	1.06	1.1	1.05	1.0
R208	765	3.73	–0.34	0.06	1.05	1.0	1.05	0.9
R209	772	3.38	0.75	0.06	1.13	2.3	1.12	2.2
R210	938	3.69	–0.05	0.06	1.38	7.2	1.37	7.0
R211	20	3.33	0.52	0.40	0.77	–0.6	0.74	–0.7
Separation: 4.29			Reliability: 0.95					
Fixed (all same) chi-square: 946.7 significance *p*: .00								

[a]589 students × 6 essays × 4 raters = 14,136 essays.

The infit and outfit mean squares in all categories are between .92 and 1.09, very close to the expected mean square value of 1.0 (Bond et al., 2020). Moreover, FACETS was used to analyze the 19 human raters to obtain the "fair average" ability measure for each essay. 3453 essays from 589 students were collected in the research. The correlation between fair average scores by human raters and IEA scores is $r = .68$.

From the calibration of the IEA categories against human ratings, approximate cut scores might be established for classroom teachers' reference. It suggests possible cutoff points for the nine categories developed for the IEA, and the relationship between the raw scores for the IEA against those human ratings. For example, a student essay with an IEA score between 37 and 62 would probably score a human rating of 3. Student essays would need to exceed an IEA measure of 83 to likely qualify for a human rating of 5, and so on.

Discussion

One persistent key problem in human essay scoring has been to do with the variability in rater leniency and severity. In the first mathematical study of rating, Edgeworth (1890)

Table 6. Calibration of IEA for student essays using human raters and IEA.

IEA	No. of essays	Observed score	Measure (Logit)	*SE*	Infit		Outfit	
					MnSq	ZStd	MnSq	ZStd
Category 1	162	2.22	–3.82	0.07	1.04	0.7	1.08	1.4
Category 2	257	2.75	–2.70	0.05	0.93	–1.6	0.93	–1.6
Category 3	441	3.02	–1.81	0.04	1.00	0.0	1.00	0.0
Category 4	499	3.27	–1.06	0.04	1.04	1.2	1.04	1.2
Category 5	670	3.63	0.05	0.03	0.93	–2.7	0.92	–2.8
Category 6	539	3.93	0.99	0.04	0.97	–0.8	0.97	–0.8
Category 7	416	4.16	1.74	0.04	1.09	2.3	1.07	2.0
Category 8	309	4.56	2.74	0.05	1.06	1.4	1.05	1.2
Category 9	160	4.99	3.87	0.07	0.99	–0.1	1.00	0.0
Separation: 49.88			Reliability: 1.00.					
Fixed (all same) chi-square: 17878.6			significance *p*: .00.					

discovered that the spread of rater leniencies was about half the spread of person abilities. While a whole gamut of descriptive statistical models have been developed to address the obvious symptoms of the problem, it was not until Linacre (1989) that a distinctly measurement foundation, based on prescriptive Rasch theory, was constructed to solve the problem. MFRM is designed to produce linear measures that adjust for rater leniency and missing data with the minimum load on the raters, while also producing useful diagnostic information about each rater's behavior. The most comprehensive current account of the benefits of Rasch measures over traditional statistical approaches was written by Eckes (2011). The unique innovation of our current research is to extend the use of MFRM to include computer-generated raw scores for essays. Not only does this provide diagnostic information that is now the default requirement in rating evaluations, the co-calibration allows us to go further than merely showing the usual human/machine agreements or correlations, and underpins the construction of AES/human score equivalences shown in Figure 5.

Although neither this, nor any AES system, can be sensitive to connotation and context, the empirical results suggest that the IEA could be used quite successfully in different assessment settings. Teachers might use AES systems to meet their students' and their own needs in an effective way (Warschauer & Grimes, 2008). In spite of the obvious limitations, the IEA could have a role in providing a positive reference index for maintaining consistency (i.e., the moderation) of grade allocations for marking English essays written in response to these NAEP prompts. Most of the prompts in both A and B versions show difficulty estimates within one logit difference which means that they are more or less equivalent in terms of the student proficiency required to obtain any particular score. Only two persuasive prompts in the 12 grade (12AP and 12BP) varied more (i.e., a difference of 1.26 logits; 2.25 and 0.99). This quite substantial difference should be investigated in future research, especially because the prompts are supposedly equivalent in terms of difficulty in writing assessment projects.

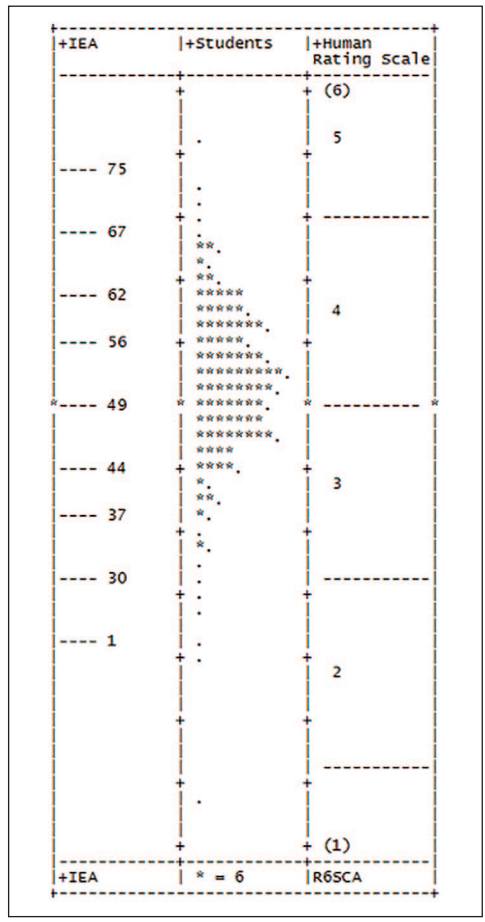

Figure 5. Conversion co-calibrations for grading student essays.

We note that the same 6-point rating scale was used in scoring all essays with the genres of narrative, informative, or persuasive texts, but each prompt has a unique scoring guide so that the level descriptors for each year level are different for each of the different prompts. Overall, a set of scales with six ratings was used to grade the essays: 1—unsatisfactory, 2—insufficient, 3—uneven, 4—sufficient, 5—skillful, and 6—excellent. From this, we conclude that the definitions of 1–6 (i.e., the rubrics) are grade-specific, even if the numerals 1–6 are not. This is similar to raw scores on a computer-adaptive test. We must know the context in which the score is made. So, any data row might be read as this: The raw score 3 (uneven) is given by Rater R206 to the essay written by Student 056 in response to Persuasive prompt A, scored against the year 4 criteria (coded as 4AP).

One might quite reasonably ask how MFRM was able to disambiguate the grade levels in these data: Does the analysis spread the student measures from the lowest

performance at the lowest grade to the highest performance at the highest grade using some sort of conventional vertical equating commonly used in multiple year-level assessments; or does the analysis spread the student measures from low performers to high performers, regardless of grade level, by relying on the highly trained raters to use the changing year-level scoring guides to make those grade-level adjustments? The quantitative empirical evidence supports the latter position.

MFRM has no statistical sleight of hand trick to disambiguate the ratings given from the essay (prompt) being rated. On each occasion, these professional raters read the essay, considered the rating criteria (rubric) for that year level, and assigned a particular rating in that specific context. (We awarded the score of 3—uneven to this rather simple essay from a year four student judged against the year 4 standard; then we awarded the same score 3—uneven, for an obviously better essay, from a Year 12 student, judged against the Year 12 rubric.) Uniquely, MFRM analyses all assessment facets—student, prompt, rating, and rater—simultaneously, in one measurement framework, so each of those facets might be considered individually. One key outcome of that analysis is evident in the spread of the difficulty estimates of the prompts in Figure 4 and the detail provided in Table 4: although the raters awarded a mean raw score rating of 3.x on 15 of the essay prompts (except for 8AP, 2.90; 12BP, 4.02; 12BN, 4.04), the MFRM Rasch measures of prompt difficulty ranged over 5 logits, from a low of −2.90 logits for 4AP to a high of +2.25 logits for prompt 12AP. This is an empirical corroboration of our expectations. While raters gave, on average, much the same raw score to all essays, the difficulties of prompts are hierarchically arranged empirically so that, in general, Level 4 prompts are easiest, then Level 8, leaving Level 12 as the most difficult prompts in this MFRM context. Similarly, the average raw rating awarded by 18 raters, aggregated across all year prompts was in the 3.x range (rater R106 was marginally more generous than the others, with the mean of ratings awarded at 4.04; see Table 5). Similar banding is revealed for student writing ability as well. Table 7 shows the average raw ratings assigned by raters at each year level, from a low of 3.51 for Year 4 to 3.85 for Year 12. However, when those raw ratings are analyzed simultaneously with the MFRM, in the context of prompt difficulty (and rater severity), those almost identical raw score ratings are spread across 2 logits—in the Rasch measure column—from −0.70 (Year 4) to +1.20 (Year 12). Those mean Rasch measures are also displayed by year level in the student distribution in Figure 4. That is, any raw score of 3—uneven is more, or less difficult to achieve, according to the context in which it was awarded, and the raters' professional use of the appropriate scoring guide for the year level of the prompt.

In this research, the best nine-level categorization for the measures of the AES system, IEA, was obtained from the collapsing of categories, following the guidelines of Linacre (1999, 2002). In other words, the optimal categorization was used in the MFRM analyses in this research, a technique that could be applied profitably in future studies comparing machine and human ratings.

The purpose of analyzing this large sub-set of the original research data (i.e., the US student essays scored by IEA and US raters who are also L1 speakers of English) is to generate MFRM analysis results of all student essays within a single language context. The analysis of the student essays in this research reveals that the IEA scoring fits the measurement requirements of the Rasch model and that no misfitting human raters are

Table 7. Average raw human ratings and mean Rasch measures by year level.

Grade	Average raw human ratings	Mean Rasch measures
4	3.51	−0.70
6	3.60	0.22
8	3.61	0.53
10	3.90	1.04
12	3.85	1.20

found. If the IEA were to be adopted as a tool for reducing the essay grading load on English language teachers in their routine school-based assessment, or low-stakes assessment, teachers could save time and effort in the actual scoring of essays, and could then spend correspondingly more time working with students individually and providing in-depth feedback on students' essays written under the usual requirements of the teacher–school essay assessment. For example, a set of, say, released NAEP prompts could be assigned to be the core essay topics for any particular level, of students in a school. The IEA might be used on a trial basis for one semester/year in a specific year level, and the proportion of essays in classroom assessment could be gradually increased. Other year levels could be involved in the trial in the next academic year.

For school-based examinations, each essay could be scored by the machine and one teacher (Rupp et al., 2019; Shermis & Burstein, 2013). The IEA-generated scores could then be used as a reference point for moderation of teacher-assigned grades across classes, so that students' essays might be re-scored by a second teacher if essays have more than some acceptable discrepancy between the IEA and teacher scores. Then a more objective and effective moderation strategy could provide more reliable comparative standard to help establish more equitable essay scoring system between teachers and across grades in a school.

Moreover, the IEA could be also used in large high-stakes assessments, although the training cost for the IEA rating engine is very high. For any new prompt, it is a pre-requisite to have at least 100–300 representative essays scored by two human raters to train the IEA system for scoring, as the IEA gives its score to an essay by comparing its content with that of the set of previously scored essays. But, this would be beyond the resources of groups of teachers, or even schools to use IEA for routine English essay marking. However, if teachers could use publicly available prompts (e.g., from NAEP, as outlined in the school-based moderation scenario, above) that have been previously submitted to the IEA for training, then obtaining IEA scores for the resultant essays could be very straight forward. It could be possible for students to have self-learning opportunities to assess their own writing of those available prompts by using the IEA. Given that the fit statistics of the MFRM for the IEA are very close to the modeled value of 1.0, the results of this research underline the recommendation that teachers, schools, and systems could rely on the IEA to score student essays at least as consistently as do specialist trained human raters. Each of these possibilities might provide insights into scoring strategies, and the challenges in achieving scoring consistency for stakeholders including students, English language teachers, raters, state

and national education authorities, as well as national testing groups in educational systems involved in English essay assessments.

Limitations and directions for future research

Further research is needed to examine the extent to which the relationships uncovered in this research are replicable. It would be useful to know whether the findings are transferable to other English language essay writing contexts, for example, where English is not the students' first language. If this research were replicated with L2 learners of English, we might expect the extent of misfit to be different. There is some existing literature on this topic (e.g., Chan & Bond, 2016), but research in this area is still very limited. It remains important to explore and evaluate the directions and theories of writing assessments that include the electronic scoring of essays, and innovations in classroom assessment of writing (Behizadeh & Engelhard, 2011).

In this research, scores of only one AES system, the IEA, were included in the data analysis. The ideal data set for large-scale research would have a number of AES measures for the essays. It seems that large-scale research including a variety of AES scores of students' essays could be conducted as a way to reveal how the various qualities of existing automated writing assessment systems impact on essay grades. For example, one AES system might be recommended as appropriate for scoring essays of a particular grade of students. The findings of such a large-scale research project must help schools or teachers choose the most suitable AES system(s) for implementing high quality, equitable scoring of essays in both low-stake and high-stake tests.

The calibration of essays and raters is improved when the network has every rater to grade every essay in the design (Lunz et al., 1990). Future research could involve raters who are L1 speakers of English scoring essays of students who are L2 speakers of English. For example, if our HK student essays were also scored by the US raters, the comparison of human raters across contexts would be well-rounded and serve as a reference guide for policy makers, educators, teachers, students, or anyone involved in the writing assessment systems in both learning English as a first or second language.

Conclusion

The connection between the educational philosophy and practices in the use of technology are of vital importance (Andrei, 2017). Although the AES systems are able to score essays similarly to human raters (Correnti et al., 2020; Wilson, 2018), they are often accepted as scoring tools that merely complement, but cannot replace expert human raters, either in high-stakes or low-stakes assessments.

In spite of significant blind spots, such as lack of responsiveness to connotation and context, the IEA could still be applied in a variety of ways in learning and testing situations. It can be utilized by teachers to accommodate students' needs and their own. If double-marking were adopted, that is, with the combination of one AES system and one human per assessment, that could be an ideal strategy for providing a more consistent essay scoring environment for students. Alongside providing consistency, the AES systems could also reduce the assessment burdens on teachers and give detailed insights to the relevant stakeholders.

Acknowledgements

We thank Mike Linacre for his valued advice concerning the structure of MFRM analyses and the interpretation of the results.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Kinnie Kin Yee Chan https://orcid.org/0000-0002-3421-6024

References

Andrei, E. (2017). Technology in teaching English language earners: The case of three middle school teachers. *TESOL Journal*, *8*(2), 409–431. https://doi.org/10.1002/tesj.280

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Banerjee, H. (2017). Test fairness in second language assessment. *Working Papers in TESOL and Applied Linguistics*, *16*(1), 54–59. https://doi.org/10.7916/D8NG62KG

Barrett, C. M. (2015). *Automated essay evaluation and the computational paradigm: Machine scoring enters the classroom* [Unpublished doctoral dissertation], University of Rhode Island. https://doi.org/10.23860/diss-barrett-catherine-2015

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*(3), 189–211. https://doi.org/10.1016/j.asw.2011.03.001

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Burstein, J., & Kaplan, R. (1996). *E-rater® scoring engine* [Computer software]. https://www.ets.org/erater/

Chan, K. K. Y. (2012). *A comparison of automated scoring engines and human raters on the assessment of English essay writing* [Unpublished doctoral dissertation], James Cook University. http://eprints.jcu.edu.au/23841/

Chan, K. K. Y., & Bond, T. G. (2016). Computer vs. human scoring in the assessment of Hong Kong students' English essays. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp. 67–88). Cambridge Scholars.

Chowdhury, T. A. (2020). Towards consistent and fair assessment practice of students' subjective writing. *International Journal of Linguistics and Translation Studies*, *1*(1), 32–41. https://doi.org/10.36892/ijlts.v1i1.14

Correnti, R., Matsumura, L. C., Wang, E., Litman, D., Rahimi, Z., & Kisa, Z. (2020). Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*, *55*(3), 493–520. https://doi.org/10.1002/rrq.281

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51. https://doi.org/10.1177/026553229000700104

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning and Assessment*, *5*(1), 49–62. https://ejournals.bc.edu/index.php/jtla/article/view/1640

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, *53*(4), 644–663. https://www.jstor.org/stable/2979547

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Erguyan, I. D., & Aksu Dunya, B. (2020). Analyzing rater severity in a freshman composition course using many facet Rasch measurement. *Language Testing in Asia*, *10*, Article 1. https://doi.org/10.1186/s40468-020-0098-3

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, *1*(2), 939–944. http://imej.wfu.edu/articles/1999/2/04/printver.asp

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2014). Implementation and application of the intelligent essay assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 66–88). Routledge.

Gottlieb, M. (2006). *Assessing English language learners: Bridges from language proficiency to academic achievement*. Corwin Press.

The Graduate Management Admission Council. (n.d.). *Analytical Writing Assessment section: Think critically and communicate your ideas*. https://www.mba.com/exams/gmat/about-the-gmat-exam/gmat-exam-structure/analytical-writing-assessment

Hamp-Lyons, L. (2019). Reflecting on the past, embracing the future. *Assessing Writing*, *42*, Article 100423. https://doi.org/10.1016/j.asw.2019.100423

Hoang, G. (2011). *Validating my access as an automated writing instructional tool for English language learners* [Unpublished master's thesis]. California State University.

Ifenthaler, D., & Dikli, S. (2015). Automated scoring of essays. In J. Spector (Ed.), *The Sage encyclopedia of educational technology* (pp. 65–68). SAGE.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated essay scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross disciplinary perspective* (pp. 87–112). Lawrence Erlbaum.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560. https://doi.org/10.1177/0265532211406422

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (1995). Misfit statistics for rating scale categories. *Rasch Measurement Transactions*, *9*(3), 450–451. https://www.rasch.org/rmt/rmt93j.htm

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103–122. http://jampress.org/JOM_V3N2.pdf

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331–345. https://doi.org/10.1207/s15324818ame0304_3

McNamara, T. (1996). *Measuring second language performance*. Longman.

McNamara, T. (2011). Applied linguistics and measurement: A dialogue. *Language Testing*, *28*(4), 435–440. https://doi.org/10.1177/0265532211413446

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing*, *29*(4), 555–576. https://doi.org/10.1177/0265532211430367

Moe, A. J. (1980). Analyzing text with computers. *Educational Technology*, *20*(7), 29–31. https://www.jstor.org/stable/44423187

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, *48*, 238–243. https://www.jstor.org/stable/20371545

Page, E. B. (2003). *Project Essay Grade (PEG®)* [Computer software]. measurementinc.com. https://www.measurementinc.com/products-services/automated-essay-scoring

Pearson. (1998). *Intelligent Essay Assessors (IEA)* [Computer software]. https://www.pearsonassessments.com

Pearson. (2019a). *General overview of WriteToLearn and its components*. https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/writetolearn/WTL-General-Overview.pdf

Pearson. (2019b). *PTE academic automated scoring* [White paper]. https://assets.ctfassets.net/yqwtwibiobs4/26s58z1YI9J4oRtv0qo3mo/88121f3d60b5f4bc2e5d175974d52951/Pearson-Test-of-English-Academic-Automated-Scoring-White-Paper-May-2018.pdf

Rotou, O., & Rupp, A. A. (2020). *Evaluations of automated scoring systems in practice*. ETS research report series. https://doi.org/10.1002/ets2.12293

Rudner, L., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer, *Practical Assessment, Research, and Evaluation*, *7*. https://doi.org/10.7275/m6xa-zg39

Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. *Journal of Technology, Learning and Assessment*, *4*(4), 86–107. https://ejournals.bc.edu/index.php/jtla/article/view/1651

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). *Automated essay scoring at scale: A case study in Switzerland and Germany* (TOEFL Research Report No. RR-86). Educational Testing Service. https://doi.org/10.1002/ets2.12249

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465–493. https://doi.org/10.1177/0265532208094273

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum.

Trace, J., Janssen, G., & Meier, V. (2017). Measuring the impact of rater negotiation in writing performance assessment. *Language Testing*, *34*(1), 3–22. https://doi.org/10.1177/0265532215594830

US Department of Education, Institute of Education Sciences. (1998). *Sample questions writing 1998*. National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pdf/main1998/1999462.pdf

US Department of Education, Institute of Education Sciences. (2002). *Sample questions: Writing 2002*. National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/pdf/main2002/2003529.pdf

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal*, *3*(1), 22–26. https://doi.org/10.1080/15544800701771580

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*(2), 1–24. https://doi.org/10.1191/1362168806lr190oa

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice*, *27*(1), 46–64. https://doi.org/10.1080/0969594X.2019.1700212

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology*, *68*, 19–37. https://doi.org/10.1016/j.jsp.2017.12.005

Wohlpart, A. J., Lindsey, C., & Rademacher, C. (2008). The reliability of computer software to score essays: Innovations in a humanities course. *Computers and Composition*, *25*(2), 203–223. https://doi.org/10.1016/j.compcom.2008.04.001

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*(3), 291–300. https://doi.org/10.1177/0265532210364643

Yan, Z., & Bond, T. G. (2011). Developing a Rasch measurement physical fitness scale for Hong Kong primary school-aged students. *Measurement in Physical Education and Exercise Science*, *15*(3), 182–203. https://doi.org/10.1080/1091367X.2011.590772