Language Testing in Asia

RESEARCH Open Access

Analyzing rater severity in a freshman composition course using many facet Rasch measurement

Inan Deniz Erguvan[1*] and Beyza Aksu Dunya[2,3]

* Correspondence: erguvan.d@gust.edu.kw
[1]Gulf University for Science and Technology, Block 5, Building 1, Mubarak Al-Abdullah Area, West Mishref, Kuwait
Full list of author information is available at the end of the article

Abstract

This study examined the rater severity of instructors using a multi-trait rubric in a freshman composition course offered in a private university in Kuwait. Use of standardized multi-trait rubrics is a recent development in this course and student feedback and anchor papers provided by instructors for each essay exam necessitated the assessment of rater effects, including severity/leniency and restriction of range in ratings among instructors. Data were collected from three instructors teaching the same course in Summer 2019, who rated the first midterm exam essays of their students and shared the scores with the researcher. Also, two students from each class were randomly selected and a total of six papers were marked by all instructors for anchoring purposes. Many-facet Rasch model (MFRM) was employed for data analysis. The results showed that although the raters used the rubric consistently during scoring across all examinees and tasks, they differed in their degree of leniency and severity, and tended to assign scores of 70 and 80 more frequently than the other scores. The study shows that composition instructors may differ in their rating behavior and this may cause dissatisfaction, creating a sense of unfairness among the students of severe instructors. The findings of this study are expected to help writing departments to monitor their inter-rater reliability and consistency in their ratings. The most practical way to achieve this is by organizing rater training workshops.

Keywords: Freshman composition, Multi-trait rubric, Many-facet Rasch model, Rater behavior, Leniency and severity

Introduction

National Council of Teachers of English (NCTE) proposes that writing is a complex skill learned over a long period of time, through a wide range of assignments, and with copious and significant feedback (Anson, Filkins, Hicks, O'Neill, Pierce, & Winn, 2013). Students must gain this complex skill in order to meet the requirements of higher education, demands of a twenty-first-century workforce, and the realization of meaningful lives. As writing has become a significant skill to master for students, developing assessment systems has also become a pressing need. In composition and writing classes, performance assessment is done through assessing student writing, either via direct or indirect methods.

There is a growing consensus in the profession that the most ideal way to survey students' composing skills is through writing, i.e. "direct" assessment. Multiple-choice testing that once dominated writing assessment is not seen adequate any longer (Barkaoui, 2011). However, direct writing assessment is also challenged, because unlike the straightforward multiple-choice assessment, the assessment of student writing, particularly in English as a second language (ESL) classes, is a challenging task for writing instructors (Huang & Foote, 2010), and there is plenty of evidence that raters from different backgrounds seem to weigh assessment criteria quite differently when they are scoring their students' essays (Barkaoui, 2010).

There is no doubt that scoring criteria play a central role in assessing examinee performance.

Thus, the scoring method should be chosen carefully before assessing student writing directly. Three methods for assessment, namely holistic, analytic and multi-trait scoring, have been widely used in writing assessment. Holistic scoring requires making an overall judgment about the quality of a student's writing, without analyzing its specific features. In addition, holistic scores provide little diagnostic information to students regarding the basis of their score or how to improve their writing (Lai, Wolfe, & Vickers, 2012). The second method, analytic scoring, involves assessment of writing through analyzing the separate components of student writing (Vacc, 1989). The third method, multi-trait scoring calls for performance evaluation under several traits. Multi-trait rubrics look like analytic rubrics because performance is evaluated in several categories in both rubric types. These terms are sometimes used interchangeably; however, the criteria in multi-trait rubrics focus on specific features of performance necessary for successful fulfillment of a given task. According to Weigle (2002), trait-based rubrics focus on a particular task and evaluate performance dimensions comparative to the requirements of that task.

The word "rubric" implies an assessment tool that describes levels of performance on a particular task and is used to assess outcomes in a variety of performance-based contexts (Hafner & Hafner, 2003). An educational rubric is a scoring device for a qualitative rating of student performance. It incorporates criteria to rate essential dimensions of performance, as well as standards of achieving those criteria (Jonsson & Svingby, 2007). The rubric tells both teachers and students what fundamental skills teachers seek for while they are assessing student performance (Arter & McTighe, 2001).

Rubrics are powerful tools that measure the performance of test takers. They provide the opportunity for reliable scoring, rather than a subjective scoring simply based on the rater's personal idiosyncrasies (Carr, 2000). Among several benefits of using rubrics, providing consistency of scoring across students, assignments, as well as among different raters is a major one. Rubrics offer a way to provide validity in assessing complex aptitudes, without forgoing the need for reliability (Kemp, Morrison, & Ross, 1998). Rubrics are also said to promote learning by making the criteria and standards clear to students and providing them with quality feedback (Arter & McTighe, 2001; Wiggins, 1998).

There are some advantages of multi-trait rubrics such as they are affiliated with the task; therefore, the teacher feedback is focused on dimensions and sub-skills

that are important in the existing learning context. Students can understand the language that they are written in, which allows them to find out more about their strengths and weaknesses (Hamp-Lyons, 1991a; McNamara, 1996; Tedick, 2002). The literature review shows that many studies have been conducted to compare holistic to analytic rubrics. Bacha (2001) found that the English as a foreign language (EFL) program would benefit from more analytic measures after comparing the holistic and analytic scores of the same texts. The researcher suggests that the holistic scoring exposed little about the performance of the students in the different components of the writing task. Researchers conclude that EFL students may have different proficiency levels in different writing components; therefore, they can benefit more from analytic scoring which provides feedback on different components of their writing skills. Brown, Glasswell, and Harland (2004) studied the reliability and validity of a New Zealand writing assessment scoring rubric and found high reliability in terms of consensus, consistency, and measurement, in spite of a short rater training. Meanwhile, Hamp-Lyons (1991a) investigated the validity of a multi-trait scoring procedure in which they claimed that the scoring method taken as a whole seemed to be highly reliable in composition assessment, and appropriate for writings from different contexts. In another study conducted by Park (2006), holistic, analytic, and multi-trait scoring methods were compared in the assessment of Korean high school students' argumentative essays. The researcher investigated the rater reliability while using the rubrics, and observed significant differences. While raters gave relatively low scores when they used holistic scoring, the inter-rater reliability was the highest in multi-trait scoring. Ghalib and Al-Hattami's research (Ghalib & Al-Hattami, 2015) investigated the performance of holistic and analytic scoring rubrics in an English undergraduate program in a university in Yemen in order to compare the students' performances using two different rubrics. According to the results, analytic scoring rubrics placed the students more accurately based on their writing ability, and were considered more reliable than holistic scoring rubrics for evaluating writing in English as a Second Language.

Essay rating is a complex and error-prone cognitive process which introduces systematic variance in performance ratings. As a result, despite the popularity of rubrics, attention should also be turned to raters themselves. Raters are central to writing performance assessment; and rater training, rater experience, and rater expertise play an important role in this process (Lim, 2011). Researchers have long recognized that rater judgments have an element of subjectivity. It is inevitable that the act of rating involves rater errors or rater biases (Myford & Wolfe, 2003), and although raters are trained to use and interpret rating scales in similar ways, rater effects also need to be studied.

Rater behavior must undoubtedly be taken into consideration in order to assess the construct in question. When raters assign scores to test takers' responses, apart from the respondents' level of performance, facets such as task difficulty, the severity of the rater, and the appropriate use of scoring rubrics that may be affecting their ratings must be taken into account (Lane & Stone, 2006). Among many potential rater errors, four major categories of rater errors have been given emphasis: (a) severity or leniency, (b) halo, (c) central tendency, and (d) restriction of range (Myford & Wolfe, 2003; Saal, Downey, & Lahey, 1980).

Severity/leniency is defined as the tendency of a rater to assign higher or lower ratings on average than those assigned by other raters, and it is commonly considered to be the most pervasive and detrimental rater effect in performance assessments (Dobria, 2011). Various factors contribute to a rater's severity or leniency including professional experience, and in some circumstances, the most experienced or senior rater may also be the most severe (Eckes, 2011).

Rater errors have been analyzed in high stake writing exams such as Test of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS). Shirazi (2019) analyzed eight raters' scores of TOEFL and IELTS essays and found that when essays were scored based on IELTS analytic rating scale, both experienced and novice raters performed quite alike. However, when the essays were holistically scored, experienced raters were more lenient and novice raters were more severe. Overall, the raters were more consistent in analytic scoring, whereas this was not the case with holistic scoring.

The aim of this paper is to examine the rating behavior of instructors while they were using multi-trait scoring rubrics in a first-year composition course (ENG 100) in a private university in Kuwait. The use of rubrics in this course is a relatively recent development. Before the implementation of online rubrics, instructors were using holistic scoring, without any previous training, standardization session, or monitoring. Multi-trait rubrics were developed 3 years ago, due to their strengths in giving detailed feedback to students that can be used for remedial action. Scoring practices within the department have become more uniform since then; however, both written and verbal student feedback tend to indicate there may be some inconsistencies in the way instructors assess their students' written performances.

First-year composition (sometimes known as freshman composition or freshman writing) is an introductory writing course in American style universities. According to the course description on Gulf University for Science and Technology website (2019), ENG 100 "serves as a bridge which facilitates students' entry into the university life by developing their high-intermediate level writing, reading, and critical thinking skills to the level required in GUST degree courses." The course assesses student learning by conducting two midterm exams and a final exam in which students are required to write five-paragraph essays within 50 min on a prompt provided by the course instructor. Student essays are assessed through a multi-trait rubric developed by the course coordinator. Course instructors upload the rubric online as embedded in Turnitin, and students receive their feedback and their score through this online system.

To find out whether instructors display rater errors, this paper tried to find answers to the following research questions:

1. Do the instructors differ in terms of their level of severity while rating the student essays? If yes, which rater is more severe/lenient than others?
2. How consistently are the instructors able to distinguish among the students in terms of their levels of proficiency?
3. Did any of the instructors' ratings show evidence of restriction of range while the instructors were using the rating scale?

Method

We used many-facet Rasch model (MFRM) to analyze the rater behavior. A *facet* is an aspect of any assessment situation that may have an influence on the measurement process. A facet can be raters, performance tasks, or examinee-related characteristics such as ethnicity, gender, etc. (Myford & Dobria, 2012). The advantage of MFRM with respect to classical approach while examining rating data is that MFRM allows an in-depth analysis of similarities and differences in ratings even when a different set of examinees are concerned. In the classical approach, interrater reliability is reported while analyzing rating data. Interrater reliability is an informative statistic, yet it is limited in detecting possible rater effects such as severity. MFRM provides a valid account of potential unwanted sources of variance in ratings such as severity/leniency or bias.

Research population

The data used for this study came from the Summer 2019 term of ENG 100. Three instructors teaching the same course during the summer school in May–July 2019 rated the first midterm exam essays of their students that took place in the first week of June 2019 and shared the results with the researchers. A total of 112 students took the course in six different sections offered by three different instructors. In regular semesters (Fall and Spring), the average number of students taking this course are around 300 to 350 students, so the sample group in Summer 2019 is around one-third of the student population taking this course every semester. Out of 112, 109 students sat for the first midterm exam and three students missed the exam for medical and personal reasons. Rater 1 had 41, rater 2 had 46, and rater 3 had 22 student essays to rate.

Around 90% of the student population in the university is Kuwaiti, and the remaining 10% of the students are from other Middle Eastern countries, such as Jordan, Syria, and Lebanon. Students' native language is Arabic; however, as the medium of instruction in the university is English and the curriculum is American, there is a strong emphasis on English writing skills throughout the university. The expected English level of students in this course is The Common European Framework of Reference for Languages (CEFR) level of B1. Therefore, we can say these students are all native speakers of Arabic with an intermediate level of English.

All three instructors are ELT teachers with around 20 years of experience in the profession and have been working in this department for around 10 years. Two of them are male, and one is a female instructor. One of the male instructors and the female instructor are native speakers of English (British and American), whereas the other male instructor is not native (Indian).

Data collection

The instructors scored the first midterm exam of the term using an online multi-trait rubric which has been in use for the past 2 years and prepared by the coordinator of the writing section. The multi-trait rubric is uploaded on Turnitin and attached to the essay writing task; therefore, scoring takes place electronically. Turnitin rubrics are accessible to students before and after scoring.

The first midterm exam requires students to write a Cause and Effect type of essay in which they either focus on the causes or effects of a global or domestic phenomenon, such as unemployment, divorce, or environmental pollution. The students are given 50 min to write a five-paragraph essay of 400–500 words. The multi-trait rubric which is used to rate essays consists of five traits with different weights:

- Introduction (20%)
- Support (body paragraphs) (40%)
- Conclusion (20%)
- Vocabulary (10%)
- Use of English and mechanics (10%)

This is a multi-trait rubric with five different traits and students can see their performance in each trait online. However, students' performances (A, B, C, D, and F) in these traits are primarily used for feedback purposes because Turnitin rubrics are designed to show only the overall score to students, rather than a clear individual score in each trait. This is a more practical method for students as they do not have to deal with individual scores for each trait and make complex calculations to see their overall score, which matters more for most students. Therefore, rather than the individual scores for each trait, the instructors were only asked to share the overall scores of their students' essays. This is why the statistical analyses were conducted based on one basic score. (See Appendix B for the rubric used for this task.)

The overall scores of students may range from 1 to 100, which is then converted to five letter grades, the highest being A (100 out of 100), B (85/100), C (75/100), D (60/100), and the lowest being F (1/100) Additional file 1.

Process

Student essays were rated by the class instructor within a week following the midterm exam, during the first week of June 2019. No special training or a norming session was provided prior to or during the rating process. The instructors were then asked to send their students' scores to the researcher. The researcher randomly selected two student essays from each section and forwarded these essays to the other two instructors for anchoring purposes. All three instructors were asked to rate the same set of six essays so that the researchers could use their ratings of those particular essays to connect the instructors and the students, thus creating a common frame of reference that would make it possible to compare all students and all instructors on the same scale.

Many facet Rasch model analyses

Many-facet Rasch model (MFRM) is a member of the family of Rasch models, that all have grounded on the basic Rasch model (Rasch, 1980). In the basic Rasch model, developed by Rasch (1980), the probability of a correct answer depends on examinee proficiency and item (or task) difficulty. The basic Rasch model formulates probability of correct response as a function of examinee proficiency and item (or task) difficulty. Thus, examinee proficiency and item difficulty are two parameters, estimated from item

response data. The estimated parameters are expressed on the same, ruler-like scale, called "*logit scale*." A logit is the measurement unit of the scale for any parameter specified in the measurement model. The higher an examinee's proficiency from an item's difficulty is the better chance of correct response the examinee has. MFRM is an extension of basic Rasch model to facilitate study of other facets of interests in assessments that typically involve human judgment. MFRM examines beyond facets of examinee and item including raters, categories, time of the measurement, etc. that may affect scores (Eckes, 2011). For a two-facet situation that we studied, the log-odds of transition from one score point to another is represented by students' proficiency parameter, instructors' severity parameter, and a threshold parameter for the rating scale. The suitable mathematical model for a two-facet model can be expressed as follows:

$$\ln\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \propto_j - \tau_k$$

Where

p_{nijk} = the probability of the essay of student n receiving a rating of k from instructor j,

p_{nijk-1} = the probability of the essay of student n receiving a rating of k-1 from instructor j,

θ_n = the writing proficiency of student n,

$\propto_j$ = the severity of instructor j,

τ_k = the difficulty of a student's essay receiving a rating of k relative to k-1.

More facets can be added to the model depending on the potential sources of variance to the scores. In the model, *severity* refers to instructors who are consistently and significantly too harsh or too lenient, as opposed to other instructors. Severity parameter is negatively oriented, meaning that the higher the severity parameter value, the lower the rating. In a rater-mediated assessment situation, persons who provide ratings are expected to perform similar levels of severity. We analyzed if the three instructors significantly differed in the level of severity they exercised when assigning ratings (research question 1). FACETS provide for each instructor a "measure" of severity in log-odd units and associated standard errors. We also looked at the rater separation index, which refers to the number of different strata of severity in the instructors. The expected value of this statistic is 0. The last index for analyzing instructors' severity is fixed (all same) chi-square and its significance. This statistic serves as a *rater homogeneity index* and tests if the instructors significantly differ in their levels of severity. When the null hypothesis of equal severity measures does not hold, pairwise comparisons of instructors' severity measures can be conducted. Lastly, we tested the difference in severity estimates of any two instructors j and k (j, $k = 1, 2, 3$) for statistical significance. *Wald statistics* (Fischer & Scheiblechner, 1970) is a commonly used index for that purpose:

$$t_{j,k} = \frac{\hat{\propto}_j - \hat{\propto}_k}{\sqrt{\left(SE_j^2 + SE_k^2\right)}}$$

Where SE_j and SE_k are the standard errors associated with severity estimates of $\hat{\propto}_j$ and $\hat{\propto}_k$ respectively. As stated by Myford and Dobria (2012), "a rater may be

generally consistent in using the rating scale but occasionally gives an unexplainable rating, given that his/her other ratings." Outfit (outlier sensitive) fit index, which indicates unexpected ratings from an instructor whose ratings are usually consistent, were used to examine if any of the instructors used the rating scale in an inconsistent manner, assigning ratings that were surprising (research question 2). If the raters are generally consistent in their rating judgments, outfit mean-square values should range between 0.7 and 1.3 for high-stakes decisions (Bond & Fox, 2015). In general, infit and outfit mean-square statistics values that fall within the range of 0.5 and 1.5 are accepted as productive for measurement (Linacre, 2002).

After analyzing the most detrimental rater effect, severity, we also examined the restriction of range effect (research question 3). Restriction of range refers to a narrower dispersion of ratings around a non-central location on the rating scale (Eckes, 2011). Overuse of certain categories may lead to overestimation of low performers' proficiency or vice versa. A useful indicator of restriction of range effect is the rater infit statistics. Unlike un-weighted mean-square indices (outfit), infit statistics are weighted by the variance of ratings, thus it is more sensitive to non-extreme unexpected ratings (Myford & Wolfe, 2003). For a low stakes assessment program, a low control limit of 0.5 is suggested for the value of infit mean square statistics (Myford & Dobria, 2012).

Results

In any MFRM analysis, the first information that is usually checked is Table 6 on FACETS (see Fig. 1, Variable Map), which displays every facet on a ruler-like variable map. The first column of the map is the logit scale on which each facet element is estimated. In the second and fourth columns, student ability and instructor severity measures are plotted along the logit scale which allows direct comparison of student writing proficiency and instructor severity. The average logit value for student proficiency is − .38 logits and the average rater severity measure is .00. The last column of the variable map provides category thresholds, appeared as horizontal lines, which are the transition points from one scale category to another. The horizontal thresholds indicate that most ratings are clustered between 70 and 80. Some categories, such as 80–86, were not distinguishable for the raters. The lower end of the scale was underutilized, signaling that we may need to reduce the number of scale points by combining lower categories. On the other hand, it is promising that instructors distinguished students well around the passing score of 70. Variable map provides quick but useful information about the rating process. The detailed analysis of instructor behavior is provided in the following sections.

Rater severity

Figure 2 (Rater Measurement Report, Table 7.2.1 on FACETS) displays important instructor-related indices including rater severity measures. Column titled "measure" displays each instructor's severity measures in log-odd units. Instructor 3 was found the most severe rater with a severity measure of .29, followed by instructor 2 with a severity measure of .24. Instructor 1 was the most lenient in ratings with

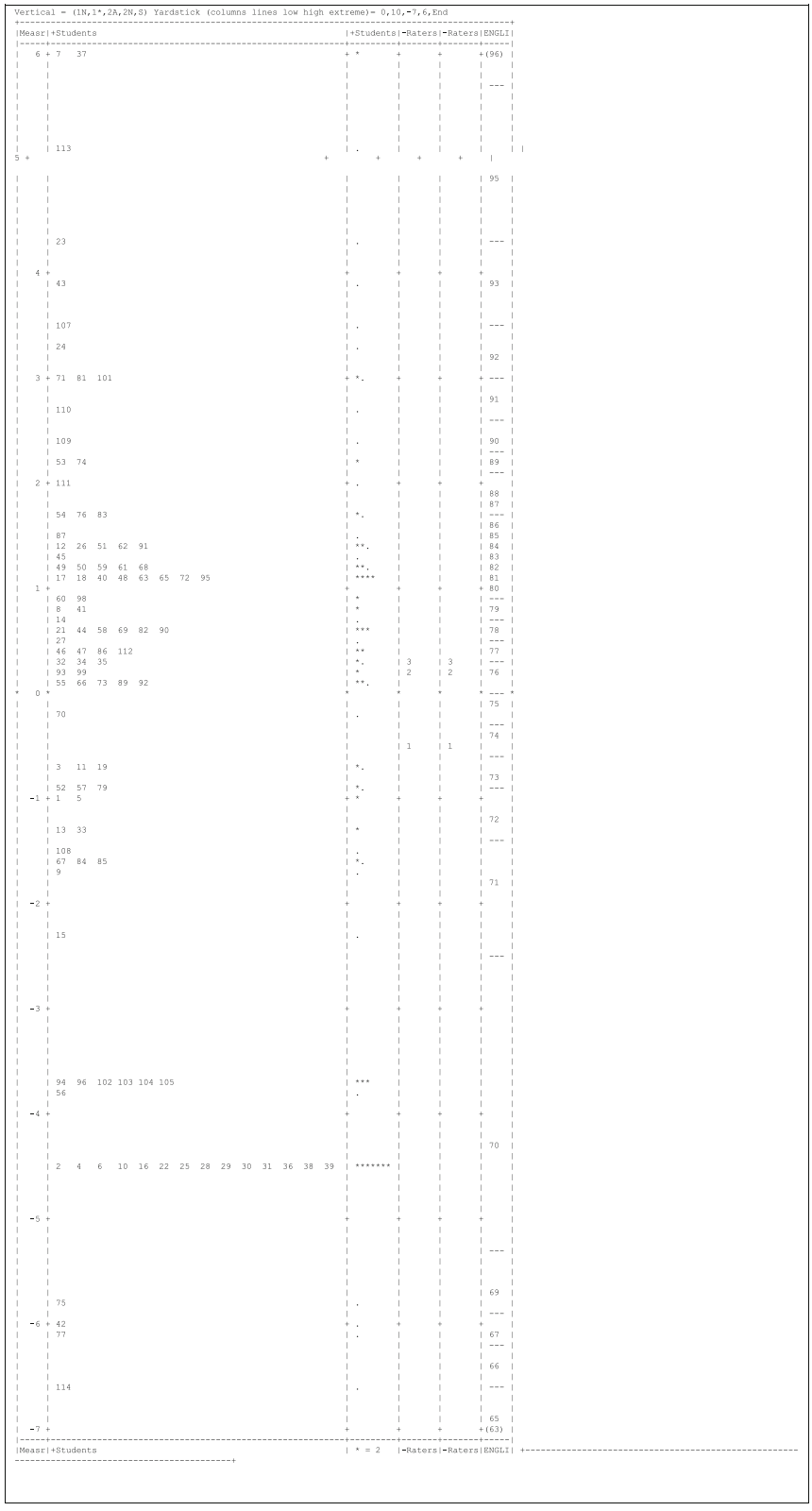

Fig. 1 Variable map (Table 6 on FACETS). Figure 1 displays every facet on a ruler-like variable map. The first column shows the logit scale on which each facet element is estimated. Second and fourth columns display student ability and rater severity measures respectively

English 100 Essay scoring 7/30/2019 1:31:08 PM
Table 7.2.1 Raters Measurement Report (arranged by mN).

Total Score	Total Count	Obsvd Average	Fair(M) Average	- Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Corr. PtBis	N Raters
2094	28	74.79	72.95	.29	.12	.77	-.5	.61	-.9	-1.69	.79	3 3
3973	52	76.40	73.09	.24	.07	.39	-3.2	.44	-3.0	1.11	.77	2 2
3611	46	78.50	75.56	-.53	.08	1.19	.7	.58	-1.3	-1.52	.52	1 1
3226.0	42.0	76.56	73.87	.00	.09	.78	-1.0	.54	-1.8		.69	Mean (Count: 3)
814.0	10.2	1.52	1.20	.38	.02	.33	1.7	.07	.9		.12	S.D. (Population)
996.9	12.5	1.86	1.47	.46	.03	.40	2.0	.09	1.1		.15	S.D. (Sample)

Model, Populn: RMSE .09 Adj (True) S.D. .36 Separation 3.90 Strata 5.53 Reliability .94
Model, Sample: RMSE .09 Adj (True) S.D. .45 Separation 4.82 Strata 6.76 Reliability .96
Model, Fixed (all same) chi-square: 61.5 d.f.: 2 significance (probability): .00
Model, Random (normal) chi-square: 1.9 d.f.: 1 significance (probability): .16

Fig. 2 Rater measurement report (Table 7.2.1 on FACETS). Figure 2 displays important rater-related indices including rater severity measures and fit statistics. Particular attention is given to the first column titled "measure" displays each rater's severity measures in log-odd units. Rater separation index indicates a number of distinct strata of severity. The fixed (all same) chi-square statistics tests if the raters significantly differ in their severity levels

a measure of − .53. In addition to the severity measures, the output also included a "fair average" for each instructor, which is the average rating for each instructor adjusted for the deviation of the instructors in that instructor's sample from the overall student mean (Myford & Wolfe, 2003). As the difference between rater fair averages' suggests on Fig. 2 (75.56–72.95 = 2.61, 75.56–73.09 = 2.47), on average, instructor 1 tended to assign ratings about 2.5 raw score points higher than instructor 2 and instructor 3. Rater separation index was found to be 6.76, indicating that there were almost seven statistically distinct levels of severity among the three instructors, which is not possible for this analysis with three raters. This finding suggests that rater separation statistics are not readily interpretable. One potential reason of this, as suggested by Myford and Wolfe (2003, p. 527), could be that the number of observations per instructor is large. It can be concluded that the spread of the rater severity measures was considerably greater than the precision of those measures (Myford & Wolfe, 2003). The fixed (all same) chi-square statistics which test if three instructors significantly differ in their severity levels, was found to be 61.5, with a significance value of less than .001. This value indicates that the severity measures for the instructors were not all the same, after allowing for measurement error. The instructors are well differentiated in terms of the levels of severity they exercised. According to Wald test statistics results, there was not a significant difference between instructors 2 and 3 in terms of severity ($t = .35$, *ns*), yet instructor 1 was significantly more lenient than both instructor 2 ($t = 5.33$, $p < .001$) and instructor 3 ($t = 5.68$, $p < .001$). The results showed that, after allowing for measurement error, the other two instructors were significantly different in terms of the levels of severity they exercised

Consistency in ratings

As seen in the Outfit MnSq column in Fig. 2, the outfit mean square values ranged from .44 to .61 for the instructors. None of the three instructors has fit values greater

than 1.3, meaning that they use the rating scale consistently across all students. Potential reasons for inconsistencies in ratings are a lack of understanding of the meaning of scale categories, thus failing to distinguish different categories; and fatigue toward the end of performance, thus not paying attention to the performance (Myford & Dobria, 2012). Our findings supported that none of these potential factors has impacted the instructors.

Rater errors (restriction of range)

According to infit values and associated standard errors, two instructors tended to rate students too consistently. Across all three instructors, infit values ranged from 0.39 to 1.19. Instructor 2's ratings yielded an infit value of 0.39 that is below the low control limit of 0.5. Instructor 3's infit value was well within the range of 0.5 to 1.5. Likewise, instructor 1's infit value was close to the expected value of 1.00. Looking at the total counts per scale point, it is observed that raters tended to use categories between 70 and 80 much more frequently than other categories. More specifically, the score point of 70, which was the cut score for passing, was the most frequently used category across all instructors.

Discussion

As research on rater training has shown, the idea of a homogeneous group of raters providing unanimous ratings is only rare, if at all achieved (Eckes, 2011; Wang & Luo, 2019). Our findings on rater severity supported that rater severity can still exist after various forms of training and experienced raters (e.g., Davis, 2016; Eckes, 2005). Despite the experience of the instructors in terms of the assessment content and teaching in the study, a significant difference in levels of severity was observed between instructor 1 and the other two instructors. As Eckes (2011) contends, "there has been a notable lack of research into the personal and situational determinants of rater severity" (p.55). Thus, we could only estimate the reasons for instructor 1's leniency at this stage. This may be resulting from instructor 1's efforts to soften their image as a tough instructor. This particular instructor has been known as a harsh instructor who does not pass students easily. Unfortunately, this has resulted in their sections closing due to a low number of registering students many times in previous years. For that reason, this instructor has been asked to make an effort to turn this negative reputation into a positive one by the Head of Department and this has probably caused this instructor to be more lenient in the summer school. The instructor used this as an opportunity to show the students he/she is not a "mean" instructor. The other two instructors do not have such a reputation in the department and are quite popular among both male and female students in the university; therefore, they may not have such concerns and restrictions while grading their students.

Despite significant severity differences among instructors, outfit and infit values that did not exceed 1.3 suggested that internal consistency was observed in each instructor's ratings. This finding was promising since previous researchers such as McNamara (1996) viewed raters' random error with respect to internal consistency more detrimental than systematic (and often explainable) rater effects. Despite instructor 1 being significantly lenient in his/her ratings, this is a manageable situation through statistical adjustments to scores in FACETS using fair average scores.

Another noticeable finding of the study was that a restriction of range in ratings around the scale point of 70 was detected. This may be resulting from a couple of factors. The first one worth emphasizing is 70, which corresponds to C–, is the minimum required score to be able to enroll in the next compulsory composition course. In the English department, there are three sequential writing courses and the lower coded course is the prerequisite of the next higher coded course. If a student gets an overall of 69 and below (an equivalent of D+, D, or F), they have to repeat the course, which costs them time and money. This cut-off score of 70 also puts pressure on the instructors to push the student up to 70, rather than leave them at 60–69 and fail the student. It should be noted that a low GPA and Pass/Fail ratio in a particular instructor's section does not reflect well on the instructor's annual merit assessment. It also gives the students the message that it is not very easy to pass the course in that instructor's course. Therefore, anything between 70 and 80 is a safe area for instructors while grading student essays with an average and slightly below the average performance.

The scoring structure of the rubric is also a reflection of this grading policy. The rubric requires the instructors to give 1 point to F, which means that particular dimension of the essay does not exist. F is used for rare cases, such as extremely poor or incomplete essays or plagiarism. Instructors assign D (60) to students for performance quite below the average. Again, D means the student is failing in that dimension and the student is not competent enough to pass the exam, and eventually the course. C, which corresponds to 75, is a safe score for an average essay and instructors may have a preference over C to D in order to raise the student's average to a score closer to C– in an effort to motivate the student and help increase the students' average. The department's overall average in major courses also hovers around a GPA of 2.5–2.7, which corresponds to 75–78. Therefore, we can say this range is also a reflection of the department and the university average.

Another possible reason for an observable preference for 70 could also stem from the relatively more relaxed atmosphere of the summer school. Both the students and instructors may have the (mis)conception that no student fails (i.e., gets a D+, D, or an F) in the summer school as summer school is supposed to be easier than the regular academic terms. Instructors may be tempted to pass below-average students with the minimum required a score of 70, i.e., C– and pass the buck to the instructor of the next course.

Conclusion and pedagogical implications

The findings of this study show that freshman composition instructors may differ in their rating behavior and this may cause dissatisfaction, creating a sense of unfairness among the students of severe instructors. Therefore, this study is expected to help this particular writing department to be more standardized in their ratings. The most common way of fulfilling this goal is training sessions, where instructors are introduced with a set of criteria and then they are asked to rate essays based on those criteria. The results show whether and to what extent they are on the same page as other raters and therefore interpret the rating criteria similarly. Fahim and Bijani (2011) suggest that rater training reduces extreme scores in terms of severity and leniency and brings them closer. Rater training aims to reduce variability and randomness of overall severity or leniency. Thus, as of Fall semester 2019, at least one norming session should be targeted a semester

among the writing instructors teaching the same course in which all instructors should rate randomly selected student essays and discuss their scoring rationale afterwards. This is supposed to reduce the severity/leniency gap among the instructors.

It should also be reminded that although there is evidence indicating the effectiveness of rater training, there is also evidence that the effects of this training may not last for a long time (Lumley & McNamara, 1995), which emphasizes the importance of regular workshops. O'Sullivan and Rignall (as cited in Shirazi, 2019), researchers in IELTS, also suggest that feedback delivered to raters systematically will probably result in more consistent and reliable examiner performance. Organizing a short standardization session before every midterm will probably be the most efficient way for departmental standardization, despite the time restrictions raters may have.

Another significant implication of this study will be on the revision of the rubric. The piling up of most students within the 70–80 range may be an indication that the rubric fails to differentiate the C– students from C and C+ students. Another band may be added to draw a line between 70 and 74 and 75–79 for a healthier and more balanced grade distribution. Also, instructors should be encouraged to assign a D to students who are clearly below the average rather than pushing them up to the C band, in order to avoid grade inflation. This revision could be based on the results of rater training workshops in which instructors should be given a chance to give feedback on the rubric and ask questions to reduce certain ambiguities the rubric may have.

The implications of this study are not limited to the particular institution or the region where the study was held. This study has implications for many rater-mediated language assessment situations, particularly in small-scale academic programs, for example, the relation between raters' demographic characteristics and various rater effects have long been researched. Specifically, raters' non-native status has attracted language assessment researchers recently (McNamara, Knoch, & Fan, 2019). The results of the study supported previous research on the effect of non-native status of raters in language assessments (Zhang & Elder, 2011). The severity issue arose in the study has not been related to the instructor's native language since one of the two instructors with similar severity measures was native (instructor 1), and other (instructor 2) was non-native. Another important message that can be generalized to other assessment situations is related to the pressure that the instructors may feel and its impact on their ratings. The cluster of the instructors' ratings around the passing score 70 may be interpreted as a result of such pressure. As previous research suggested (Goodwin, 2016), at important cut points, additional trainings may be beneficial for raters not only to facilitate the use of the scale but also to reduce the pressure they feel.

Limitations and suggestions for the further research

The limitations of this study also suggest several directions for further research on writing assessment in freshman composition and any other writing courses. First of all, this study was conducted during the summer school over a relatively small number of instructors and student population. In regular semesters (Fall and Spring), the number of instructors teaching Eng 100 is around five and six and the number of students is approximately 300–350. Thus, this study should also be conducted again in the same course with a higher number of participants during the

regular academic terms. Also, a similar method for measuring instructors' rating behavior should be implemented in the successive writing courses offered by the department. Higher numbers are expected to contribute to the reliability and consistency in grading in all writing courses and improve the quality of assessment in any department that offers writing courses.

Another limitation was the genre that was assessed by instructors. An additional study on rater behavior in assessing different genres, using a genre-specific multiple-trait scoring rubric, should display how instructors are using the multiple-trait scoring rubric across various genres. Also, investigating the rating process through think-aloud protocols and interviews with instructors will be able to show more information about what is going on in the instructor's mind.

Appendix A

Table 1 Student scores sent by raters

Student no	Midterm 1
Rater 1	
Student 1	74
Student 2	70
Student 3	75
Student 4	70
Student 5	74
Student 6	70
Student 7	96
Student 8	83
Student 9	72
Student 10	70
Student 11	75
Student 12	88
Student 13	73
Student 14	82
Student 15	71
Student 16	70
Student 17	86
Student 18	86
Student 19	75
Student 20	97
Student 21	81
Student 22	70
Student 23	95
Student 24	93
Student 25	70
Student 26	88
Student 27	80
Student 28	70

Table 1 Student scores sent by raters *(Continued)*

Student no	Midterm 1
Student 29	70
Student 30	70
Student 31	70
Student 32	79
Student 33	73
Student 34	79
Student 35	79
Student 36	70
Student 37	96
Student 38	70
Student 39	70
Student 40	86
(Rater 2)	
Student 1	79
Student 2	80
Student 3	80
Student 4	81
Student 5	72
Student 6	88
Student 7	85
Student 8	75
Student 9	70
Student 10	72
Student 11	77
Student 12	80
Student 13	78
Student 14	80
Student 15	81
Student 16	79
Student 17	58
Student 18	79
Student 19	75
Student 20	71
Student 21	80
Student 22	77
Student 23	74
Student 24	91
Student 25	79
Student 26	75
Student 27	88
Student 28	67
Student 29	85
Student 30	66
Student 31	55

Table 1 Student scores sent by raters *(Continued)*

Student no	Midterm 1
Student 32	72
Student 33	60
Student 34	91
Student 35	77
Student 36	84
Student 37	71
Student 38	71
Student 39	76
Student 40	82
Student 41	60
Student 42	75
Student 43	77
Student 44	81
Student 45	75
Rater 3	
Student 1	70
Student 2	79
Student 3	70
Student 4	20
Student 5	78
Student 6	75
Student 7	60
Student 8	91
Student 9	70
Student 10	70
Student 11	70
Student 12	70
Student 13	60
Student 14	92
Student 15	71
Student 16	89
Student 17	90
Student 18	87
Student 19	76
Student 20	95
Student 21	65

Appendix B

Table 2 ENG 100 rubric

	A 100	B 85	C 75	D 60	F 1
Introduction (20%)	The hook is catchy and there is adequate info to provide background for the topic. There is a well-focused thesis statement that introduces the essay and clearly addresses all elements of the writing prompt.	The hook exists but may not be very creative; the background info could have been more in-depth. Thesis statement introduces the topic and the opinion of the author.	At least 1 introduction element is lacking or irrelevant. Readers can identify the purpose of the essay in the thesis, but it is not very clear.	At least 2 elements of the introduction are missing or totally off-topic. Thesis does not introduce the topic.	There is no introduction paragraph or none of the introduction elements is present.
Support (Body paragraphs) (40%)	Each body paragraph contains a topic sentence and provides relevant details. Body paragraphs contain a well-developed explanation, analysis, and discussion that demonstrate understanding.	Each body paragraph contains a topic sentence that is adequately supported by relevant concrete details. Essay contains some explanation, analysis and discussion that show understanding.	Topic sentence is stated in some body paragraphs. Supporting details are relevant, but some key issues are unsupported. Support for the thesis statement is weak. Essay does not contain enough explanation or analysis.	Topic sentences are weak. Supporting details are unclear or not related to the topic. Thesis statement lacks proof. Essay contains explanation, analysis or discussion that is not correct/off-topic.	No topic sentence OR no concrete details are present. The paragraph contains no facts, details, or examples. No attempt to explain, or analyze the information it presents.
Conclusion (20%)	Conclusion is effective and gives readers a sense of closure; restates the thesis and gives a creative personal opinion.	Conclusion provides a sense of closure but personal opinion is not creative, or not restated adequately.	The conclusion is logical, but it does not provide closure for the essay. No creative personal opinion.	Conclusion exists but it brings up a new topic or does not sum up the essay effectively. It lacks personal opinion.	There is no clear conclusion, the paper just ends.
Vocabulary (including transitions) (10%)	Uses a range of academic words and phrases accurately. The choice and placement of words is accurate. Transitions vary and are used appropriately.	There is evidence of academic vocabulary as well as correct transitions. The choice and placement of words are mostly accurate, 3–4 minor errors may exist.	There is some evidence of academic vocabulary, but the writing lacks variety. Transitions may be awkward in 1–2 paragraphs. There may be 5–6 word choice errors.	Academic vocabulary is limited. The author makes 7–10 errors in word choice that interfere with understanding. Transitions are incorrect in 3–4 paragraphs.	There is no evidence of academic vocabulary. Errors in word choice make sentences unreadable (more than 10). Transitions do not exist or are not used correctly.
Use of English and Mechanics (10%)	All sentences are well constructed and have varied structure. Punctuation is used accurately. The author makes no or 1–2 minor errors in grammar, very few typos	Most sentences are well constructed, except for 3–4 minor errors, and have varied structure. Punctuation and spelling are mostly	The majority of sentences are simple. Max 5 sentences may be ill-constructed. The author makes max 5 errors in grammar, punctuation and spelling that	More than 5 sentences sound awkward, are repetitive, or are difficult to understand. Spelling and punctuation	The essay is unreadable because of errors in grammar (more than 10). Errors in spelling, capitalization or punctuation make the text

Table 2 ENG 100 rubric *(Continued)*

A 100	B 85	C 75	D 60	F 1
in spelling.	accurate (max 3 errors). The author makes no more than 3–4 errors in grammar and mechanics, and they do not interfere with understanding.	may interfere with understanding.	errors (max 10) impede understanding.	difficult to read.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40468-020-0098-3.

Additional file 1. Dataset for the study.

Abbreviations

EFL: English as a foreign language; ESL: English as a second language; MFRM: Many-facet Rasch model

Acknowledgments

We confirm that there are no other persons who satisfied the criteria for authorship but are not listed. No other person has contributed towards this manuscript.

Authors' contributions

DE wrote the literature review, collected the data, and interpreted the findings of the data analysis. BAD analyzed the raw data by MFRM and prepared the figures. We confirm that the manuscript has been read and approved by all named authors. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We understand that the Corresponding Author (DE) is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). She is responsible for communicating with the other authors about progress, submissions of revisions, and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author.

Authors' information

DE holds a PhD in Educational Administration from Marmara University in Turkey. She taught English as a second language in various universities in Turkey for 12 years and since 2011 she has been teaching academic writing skills as an Assistant Professor in Gulf University for Science and Technology in Kuwait.
BAD holds a Ph.D. in Educational Psychology, specialized in Measurement and Statistics from University of Illinois at Chicago (UIC) and a M.Ed.from Boston College. She has worked as a Visiting Clinical Assistant Professor/Assessment Coordinator at UIC between 2017 and 2019. She works as an Assistant Professor at Bartin University, Turkey.

Availability of data and materials

Data set is available as a supplementary file submitted in the system.

Competing interests

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no financial support for this work that could have influenced its outcome.

Author details

[1]Gulf University for Science and Technology, Block 5, Building 1, Mubarak Al-Abdullah Area, West Mishref, Kuwait. [2]University of Illinois at Chicago, Chicago, IL, USA. [3]Bartin University, Bartin, Turkey.

Received: 22 October 2019 Accepted: 9 January 2020
Published online: 05 February 2020

References

Anson, C., Filkins, S., Hicks, T., O'Neill, P., Pierce, K.M., Winn, M. (2013). National Council of Teachers of English Position Statement on Machine Scoring. Retrieved from http://www2.ncte.org/statement/machine_scoring/

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. California: Corwin Press.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29*, 371–383. https://doi.org/10.1016/S0346-251X(01)00025-2.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly, 44*(1), 31–57. https://doi.org/10.5054/tq.2010.214047.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279–293. https://doi.org/10.1080/0969594X.2010.526585.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York: Routledge.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing., 9*, 105–121. https://doi.org/10.1016/j.asw.2004.07.001.

Carr, N. T. (2000). A comparison of the effects of analytic and holistic rating scale types in the context of composition tests. *Issues in Applied Linguistics., 11*(2), 207–241 Retrieved from https://escholarship.org/uc/item/4dw4z8rt.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117–135.

Dobria, L. (2011). Longitudinal Rater Modeling with Splines. Retrieved from ProQuest digital dissertations. (UMI number: 3472389).

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197–221.

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

Fahim, M., & Bijani, H. (2011). The effect of rater training on raters' severity and bias in second language assessment. *Iranian Journal of Language Testing, 1*(1), 1–16 Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.457.2858&rep=rep1&type=pdf.

Fischer, G. H., & Scheiblechner, H. H. (1970). Algorithms and programs for Rasch's probabilistic test model. *Psychologische Beitrage, 12*, 23–51.

Ghalib, T., & Al-Hattami, A. (2015). Holistic versus analytic evaluation of EFL writing: A case study. *English Language Teaching, 8*, 225–236. https://doi.org/10.5539/elt.v8n7p225.

Goodwin, S. (2016) A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. Assessing Writing 30:21-31.

Gulf University for Science and Technology (2019). Course descriptions: College of Arts and Sciences Courses. Retrieved from https://www.gust.edu.kw/content/college_arts_and_sciences_courses

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509–1528. https://doi.org/10.1080/0950069022000038268.

Hamp-Lyons, L. (1991a). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Norwood: Multilingual Matters.

Huang, J., & Foote, C. (2010). Grading between lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly, 7*(3), 219–233. https://doi.org/10.1080/15434300903540894.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002.

Kemp, J. E., Morrison, G. R., & Ross, S. M. (1998). *Designing effective instruction* (2nd ed.). Upper Saddle River: Prentice Hall.

Lai, E. R., Wolfe, E. W., & Vickers, D. H. (2012). *Halo effects and analytic scoring: A summary of two empirical studies research report*. New York: Pearson research and innovation network Retrieved from https://images.pearsonassessments.com/images/tmrs/HaloEffects_and_Analytic_scoring.Pdf.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 387–431). Wesport: ACE/Praeger.

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. (2002). What do infit and outfit, mean-square and standardized mean? Rasch Measurement Transactions, 16 (2), 878.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71. https://doi.org/10.1177/026553229501200104.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment*. Oxford: Oxford University Press.

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

Myford, C. M., & Dobria, L. (2012). *FACETS introductory workshop tutorial*. Chicago: University of Illinois.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement. *Journal of Applied Measurement, 5*(2), 189–223.

Park, J. (2006). The study on the rater reliability of three scoring methods in assessing argumentative essays: Holistic, analytic, and multiple-trait scoring methods. Retrieved from http://sspace.snu.ac.kr/bitstream/10371/95933/1/The_Study_on_the_Rater_Reliability_of_Three_Scorin. pdf.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press (Original work published 1960).

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413–428.

Shirazi, M. A. (2019). For a greater good: Bias analysis in writing assessment. *SAGE Open, 9*(1), 1–14. https://doi.org/10.1177/2158244018822377.

Tedick, D. J. (2002). Proficiency-oriented language instruction and assessment: A curriculum handbook for teachers. In *CARLA working paper series*. Minneapolis: University of Minnesota, the Center for Advanced Research on Language Acquisition.

Vacc, N. N. (1989). Writing evaluation: Examining four teachers' holistic and analytic scores. *The Elementary School Journal, 90*(1), 87–95.

Wang, J., & Luo, K. (2019). Evaluating rater judgments on ETIC advanced writing tasks: An application of generalizability theory and many-facets Rasch model. *Papers in Language Testing and Assessment, 8*(2), 91–116.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31–50.

Publisher's Note