Psychological Test and Assessment Modeling, Volume 63, 2021 (3), 273-304

The Diagnostic Rating System: Rater Behavior for an Alternative Performance Assessment Rating Method

Nnamdi C. Ezike[1], *Allison J. Ames*[2]

Abstract

The Diagnostic Rating System (DRS), a novel system for rating performance assessments, purports to reduce rater cognitive load when applying traditional rubrics. This logic-based rating system, developed consistent with expert cognitive processes during performance assessment evaluation, asks a series of explicit questions to essay raters. We applied the DRS approach to an established assessment protocol in the context of ethical reasoning. The fully crossed rating study had 12 raters, each rating 30 student essays. Each rater rated each essay twice: once using a DRS and once with a traditional rubric, with the rating method counter-balanced so that half of the raters used the DRS first and half used the traditional rubric first. Many-facet Rasch measurement equating methods revealed that the raters vary in their severity levels on both rating methods. Overall, the findings suggest that the two rating methods are comparable. Correlation between the estimated examinee proficiency levels on the DRS and rubric was high. Novice and expert raters show high levels of consistency on the rubric, but novice raters were more consistent in ratings on the DRS.

Keywords: Diagnostic Rating System, performance assessment, rubric, raters, many-facet Rasch measurement

[1] *Correspondence concerning this article should be addressed to:* Nnamdi C. Ezike, M.S.
Ph.D. Candidate in Educational Statistics and Research Methods,University of Arkansas
257 Graduate Education Building,751 W Maple Street,Fayetteville, AR 72701,ncezike@uark.edu

[2] University of Arkansas

Performance assessment requires an examinee to demonstrate or perform a task rather than respond to a selected response item format, and offer a "true-to-life format" for the measurement of assessment outcomes (Leighton, 2019). Because the performance assessment response space is more complex than selected response items (Ames & Luecht, 2018), scoring of performance assessments is also more complex, relying on human raters or computer-automated processes to judge the quality of tasks assigned to examinees. Whether computer-automated or human-scored, an accurate and reliable judgment requires the use of well-defined, transparent scoring criteria, such as those articulated in a rubric. Consistent and correct rubric application may result in scores that accurately reflect examinees' abilities (Downing, 2006). However, traditional rubrics can have a high rater cognitive load because raters must interpret the rubric to make score-based interpretation (Bejar, 2012; Wolfe, 1997). For example, raters must interpret the scoring criteria, as well as the quality of the assessment product in relation to the criteria. Because of the interpretive judgments, it is important to determine whether raters are interpreting the rubric in the same way. Low inter-rater reliability (i.e., poor agreement among raters) and low intra-rater reliability (i.e., poor agreement with a rater regarding their own ratings) remain concerns for rubric use in scoring performance assessments.

A potential solution to mitigate the concerns of human judgments is the use of a computer-automated system. Automated essay scoring (AES) systems have received considerable attention in the last two decades. The AES systems are based on machine-learning techniques using scores obtained from human raters. Therefore, the success of the AES depends on how well the systems are calibrated and the quality of the human-scored essays used to"train" the systems. Wind, Wolfe, Engelhard, Foltz, and Rosenstein (2018) describe the AES as a scoring model that consists of a set of textual features obtained from the essays and one or more scoring algorithms which have parameters that associate the features to the scores for the essays. One begging question is whether AES can substitute human raters. Previous studies have compared the agreement between computer-automated and human-scored ratings. Attali, Bridgement, and Trapani (2010) found that scores on the computer-automated and human raters were comparable. Similarly, Cohen, Levi, and Ben-Simon (2018) and Weigle (2011) found that the AES was more consistent than human ratings. Although, Cohen et al. (2018) further stressed that the scores obtained using the AES were less valid. Perhaps this points to one of the weaknesses of the AES. The AES system cannot directly assess some of the more cognitively demanding aspects of writing proficiency, such as audience awareness, argumentation, critical thinking, and creativity (Zhang, 2013). This calls for a need for a scoring system that combines human raters and automated scoring system. This article describes the Diagnostic Rating System (DRS), an automated rating system modeled to guide the cognitive processes raters employ to assign scores as an alternative, or complementary, approach to a more traditional rubric approach. The DRS purports to lessen the cognitive load burden placed on human raters when typically scoring using traditional rubrics. Many-facet Rasch measurement (MFRM; Linacre, 1989) methodology is used to compare the DRS approach with a traditional rubric-based scoring approach.

1. Background: Scoring Performance Assessments

Numerous studies have cataloged the differential rater behavior that contributes to measurement variation in ratings. Potential sources of bias include, (a) central tendency effect, (b) halo effect, (c) restriction of range, and (d) severity or leniency (Engelhard, 1994; Myford & Wolfe; 2003; Saal, Downey, & Lahey, 1980). Underlying many of these concerns are the subjective interpretation by raters of rubric scoring processes (Johnson, Penny, & Gordon, 2009; Linn, Baker, & Dunbar, 1991). Popham (1990) asserts that the rating procedure could introduce bias in performance assessment. Rating too many traits at a time, fatigue due to long hours of rating, and the use of "new online rating procedure[s]" may introduce errors in the rating operation (Myford & Wolfe, 2003).

The quality and training of raters are essential for achieving higher consistency of scores among raters, which is an important factor needed to improve validity evidence and inter-rater reliability. Because human raters differ in their understanding of the rating scales, methods, and operations of the rating task, some background work must be completed before rating begins. Training of raters is an integral aspect of the rating process because qualified raters are important to "achieving and maintaining a high degree of consistency and reliability in the scoring of examinee's responses" (Johnson, Penny, & Gordon, 2009). Although an important part of the scoring process, previous research has shown that rater training has not proven to be effective in significantly reducing measurement variation among raters (Weigle, 1998). Welch (2006) suggests that higher levels of reliability can be achieved with "well-articulated scoring rubrics and well-defined and monitored scoring processes."

Despite potential shortcomings, there are several benefits of using human raters. Zhang (2013) pointed out three strengths of using human raters to score performance assessment: (a) the use of cognitive processes to evaluate the information provided by examinees, (b) connecting with their prior knowledge to make decisions, (c) can make judgment on the quality of an examinee work based on their understanding of the content. Using e-rater in conjunction with human raters, Enright and Quinlan (2010) note that human raters provide more feedback especially on the quality of ideas and content. They also found evidence of validity when e-raters are complemented by human raters (Enright & Quinlan, 2010). Thus, additional human cognitive processes by human raters are expected to give a more accurate representation of the proficiency of the examinees.

It is for this reason that alternative, or complementary, approaches to traditional rating procedures are still needed. The DRS is one attempt at addressing rater effects such as inter-rater reliability, restriction of range, and central tendency issues of rater-mediated assessments.

1.1 Diagnostic Rating System

The quality of an examinee's work is mainly judged using analytic or holistic rubrics. Holistic scoring is based on the overall assessment about the quality of the work of an examinee, whereas, in analytic scoring, the quality of an examinee's work is judged based on the different dimensions of the tasks (Eckes, 2011; Jonsson & Svingby, 2007). As East (2009) summarizes, there are several lexical features of essays that impact scoring on holistic rubrics, and Knoch (2011) says, "raters may read holistically and adjust analytic scores to match holistic impression." Thus, despite specific scoring criteria to guide raters, essay raters may use their own standards to assign scores. One solution may be to apply the comparative judgment approach (Pollitt, 2004; Thurstone, 1927) which has been argued to result in comparable reliability as traditional rating methods with reduced training and is less cognitively demanding for raters (Steedle & Ferrara, 2016). In the comparative judgment approach, raters make pairwise judgments about the quality of examinees' work by choosing the better of two essays (Sims et al., 2020). Another solution is to remove the direct assignment of scores away from raters and present more objective, explicit questions to raters. This can be achieved using the DRS.

A DRS is "a logic-based rating system modeled on expert cognitive processes during evaluation of performance assessment" (Curtis & Ames, 2018). When assigning scores to a performance assessment, raters using a rubric respond to a set of implicit questions (e.g., "does the artifact meet the criteria description for this score level?"). A DRS makes these questions explicit by asking a series of selected-response questions, as opposed to the questions being implicit with the traditional rubric approach. This may seem similar to a checklist at first, but the difference is two-fold. The DRS is automated, such that only relevant explicit questions are asked of raters and the raters do not directly assign scores. However, the DRS assigns the scores based on the questions answered by the raters.

As a tool to fully describe the differences in the DRS when compared to a rubric or checklist, we present the DRS in the context of a mid-Atlantic university's (*the University*) construct of ethical reasoning. At the University, the construct is operationalized as an active decision-making process when confronting an ethical dilemma. Engaging in ethical reasoning requires viewing the dilemma through different perspectives, and asking oneself multiple, open-ended considerations, framed as Eight Essential Questions (8EQ; Ethical Reasoning in Action, n.d.):

1. Fairness—How can I act equitably and balance legitimate interests?
2. Outcomes—What achieves the best short- and long-term outcomes for me and all others?
3. Responsibilities—What duties and/or obligations apply?
4. Character—What action best reflects who I am and the person I want to become?
5. Liberty—How does respect for freedom, personal autonomy, or consent apply?
6. Empathy—What would I do if I cared deeply about those involved?

7. Authority—What do legitimate authorities expect of me?
8. Rights—What rights apply?

This study focuses on a student learning outcome that states "students evaluate courses of action by applying the 8EQ framework in their own personal, professional, and civic ethical dilemmas." The Ethical Reasoning – Writing (ER-WR) essay assessment is a measure of the outcome, traditionally evaluated using the ER-WR Rubric (see Figure 1; James Madison University, 2014).

When scoring student essays, raters using the ER-WR Rubric scoring Element A (Ethical Situation: Identifying ethical issue in its context) should respond to the implicit questions, "*Does the student identify an ethical issue in its context in the essay?*" If there is an ethical issue identified, raters must then ask themselves, "*Is the reference to decision option implicit, and/or is there little context offered to the decision options?*" and so on as the rater progresses through the scoring criteria. As earlier noted, a rater may still be idiosyncratic despite a well-developed rubric. Our ER-WR DRS makes the ideal implicit questions, and possible answers, available to raters as clear questions. Raters simply answer the explicit questions without seeing scoring criteria or numeric score categories. In this way, the DRS is akin to an automated decision tree or flow chart, used to guide performance assessment ratings (Figure 2). Unlike a checklist, when scoring with the DRS, raters do not assign scores at all. The process is automated based on rater responses to the DRS questions. For example, consider the ER-WR Rubric's Element A. If the examinee made no reference to decision options, rubric raters would assign a score of zero. However, with the ER-WR DRS, raters reply directly to the question "*Did the author reference any decision options?*" If the response is "*no*," then the DRS automated scoring assigns a score of 0 for Element A. If the response is "*yes*," then the DRS collects more information before a score can be assigned. The DRS scoring process continues with these dichotomous decision tree questions until a terminal node, the score, is reached. A DRS maps and guides the rater cognitive processes used to assign scores by asking them to respond directly to specific and relatively unambiguous questions. Each rater receives the same set of explicit questions at the beginning, but may differ at the end depending on the answers to the previous questions provided by the raters. Ideally, this process removes any possibility of adjusting scores based on a holistic reading of the essay.

Insufficient 0	Marginal 1	Good 2	Excellent 3	Extraordinary 4
A. Ethical Situation: Identifying ethical issue in its context				
No reference to decision option(s).	Implicit reference to decision options AND/OR little context given regarding decision option(s).	Explicit but unorganized reference to decision option(s) and context.	Clear description of decision option(s) and context.	Meets criteria for *Excellent* AND… • Context treated with nuance • Builds tension with organization and word choice.
B. Key Question Reference: Mentioning the 8 KQs or equivalent terms				
Reference to zero or only one key question.	Vague references to key questions OR only two key questions referenced.	References four key questions.	References six key questions.	References all eight key questions.
C. Key Question Applicability: Describing which of the 8 KQs are applicable or not applicable to the situation and why				
No rationale provided for the applicability or inapplicability of any KQs to the ethical situation.	Provides a rationale for the applicability or inapplicability of two key questions to the ethical situation.	Provides a rationale for the applicability or inapplicability of four key questions to the ethical situation.	Provides a rationale for the applicability or inapplicability of six key questions to the ethical situation.	For all eight questions provides a rationale for its applicability or inapplicability to the ethical situation.
SPECIAL NOTE: If author identifies fewer than three applicable KQs, then Criteria "D" and "E" can be scored no higher than (1) "Marginal"				
D. Ethical Reasoning: Analyzing individual KQs				
No attempt to analyze any of the referenced key questions.	Analysis attempted using two or more key questions. Typically incorrect ascription of the key questions to the ethical situation. Account is unclear, disorganized, or inaccurate.	Analysis attempted using three or more key questions. Basically accurate ascription of the key questions to the ethical situation. Account is unclear or disorganized.	Analysis attempted using three or more key questions. Accurate ascription of the key questions to the ethical situation. Account is clear and organized.	Meets criteria for *Excellent* AND… Nuanced treatment of key questions, for example: • elucidates subtle distinctions • uses analogies or metaphors • considers different issues within same key question.
SPECIAL NOTE: If Criterion "D" is scored a 0 or 1 then Criterion "E" can be scored no higher than (1) "Marginal"				
E. Ethical Reasoning: Weighing the relevant factors and deciding				
No judgment is presented OR judgment presented with no rationale.	Uses products of the analysis and provides some weighing to make a decision. Account is unclear, disorganized, or inaccurate.	Conveys weighing approach using analysis products. Provides an intelligible basis for judgment.	Meets criteria for *Good* AND…. Logically terminates in decision that will be reached.	Meets criteria for *Excellent* AND… Products of analysis weighed to make judgment compelling.

Figure 1. ER-WR Rubric (James Madison University, 2014)

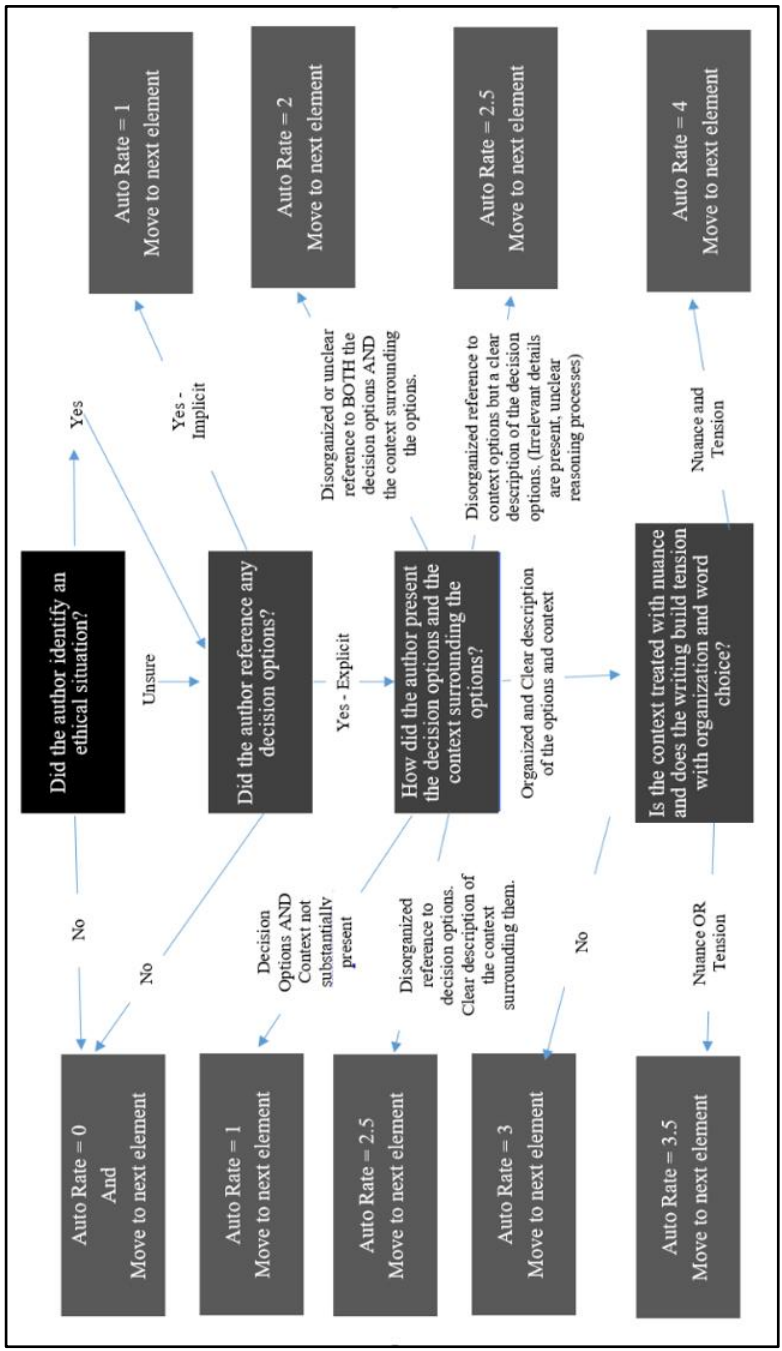

Figure 2. Element A Logic Map of ER-WR DRS (Curtis & Ames, 2018)

Research Questions

To evaluate performance of the ER-WR DRS, this study examines the following research questions:

1. Are rater effects (such as severity or leniency) mitigated with the use of the DRS method of scoring ER-WR essays?
2. How does examinees' proficiency compare across rating methods?
3. Does rater experience impact the quality of ratings across rating methods?
4. How did elements on the rating scale function across rating methods?

2.0 Method

2.1 Assessment and protocol

The ER-WR performance assessment instrument consists of an essay prompt and scoring criteria rubric. For the assessment, examinees have 55 minutes to compose an essay of at least 250 words, without additional 8EQ resources. The ER-WR prompt is:

> *"Often in life, we encounter situations that are ethically complicated. For example, if you saw a hungry child steal fruit from a grocery store, you'd likely think of many reasons to report the person and many reasons not to do so. The faculty and staff at [the University] are interested in the ethical reasoning thought process in which students engage when confronted with such situations. For this assessment, please…*
>
> *(1) Explain a complicated, ethically significant choice you faced: a choice that required a lot of thinking and deliberation.*
>
> *(2) Indicate the ethical considerations that you deem relevant to this particular situation and why, as well as which ones are not relevant and why.*
>
> *(3) Be sure to clarify your ethical reasoning process as much as possible. Try to provide an ethical analysis that is as rich and multifaceted as possible.*
>
> *(4) Lastly, be sure to say what decision you made and why."*

Each participant has engaged in at least one university-wide intervention during freshman orientation. There are multiple opportunities for 8EQ exposure through course-specific interventions, residence hall programming, campus engagement on social media, and academic integrity initiatives, although only the 90-minute freshman orientation intervention is mandatory. Results are used at the university-level to evaluate the effectiveness of ethical reasoning programming. Students are asked in a separate questionnaire about their exposure to the 8EQ through courses, residence halls, and student

affairs programs. There has been significant research into the validity evidence of the rubric and program effectiveness (e.g., Linder et al., 2020; Sanchez et al., 2017).

2.2 Pilot study

A 2017 pilot study suggested that raters rate differently when using the ER-WR DRS compared to the ER-WR Rubric (Curtis & Ames, 2018). Quantitatively, pilot study average scores and generalizability theory reliability coefficients were higher for the DRS method than the traditional rubric method. Qualitatively, experienced raters reported higher levels of critical thinking about which 8EQ are present and analyzed in students' essays when rating with the DRS. However, the initial pilot study had design limitations. Specifically, experienced raters rated using a traditional rubric and novice raters used the DRS so that the effects of rating method were confounded with the effects of rater experience. To address these concerns, the current study raters used both the DRS and rubric scoring methods to facilitate a comparison of rating processes across rating methods.

2.3 Participants

In the present study, we randomly sampled first-semester students during a University-wide assessment day (fall semester 2016) and then re-assessed the same students after four semesters (spring semester 2018). Of this sample, we chose 30 essays, comprising a mixture of first- and fourth-semester student responses. There were twelve raters, consisting of faculty (n=8), staff (n=2), and graduate students (n=2) at the University. Specifically, there were six raters with rating experience using the ER-WR rubric and six raters who had not rated with the ER-WR rubric. No raters had DRS rating experience prior to the experiment. All raters participated in a half-day core workshop that describes of the origins and application of the 8EQ process in order to become raters and teachers of 8EQ artifacts. In addition, raters are involved in some way with the 8EQ on campus, as teachers of the 8EQ approach, facilitators of a campus-wide intervention using the 8EQ approach, researchers of ethical reasoning in their discipline-specific context, or assessment specialists involved in the development and analysis of assessments designed to measure knowledge and application of the 8EQ. Seven raters were female and five were male.

Each rater assigned ratings to all 30 essays using the ER-WR Rubric and ER-WR DRS. To mitigate any potential rater "carry-over" effect between rating methods, the order of ratings was randomly assigned to raters – on day one, all raters attended a one-hour review workshop on the 8EQ approach. After the 8EQ workshop, the 12 raters were divided into two groups: one-half of the raters received training on the rubric while the other half received training on the DRS. Because of the nature of the DRS rating approach, raters using the DRS do not assign scores, whereas they do assign scores using the ER-WR rubric approach. Thus, we felt that carry-over would be minimal for raters.

Each training lasted approximately two hours. The training session involved a brief orientation to the scoring system and three sample essays that were scored by raters using a think-pair-share approach. Raters read the essays, scored them, discussed with a partner about why specific ratings were assigned, and then discussed with the entire rating group. Facilitators, experts in the 8EQ content and familiar with the assessment, provided feedback on the ratings and pointed out aspects of the essays that indicated levels of scoring criteria for the rubric and affirmative to different questions in the DRS approach.

Raters spent the rest of the day rating 30 essays using the method on which they were trained. On the second day, raters trained on the method that they did not participate in the previous day. The second-day training lasted approximately two hours following the same training format. Raters then spent the rest of the second afternoon rating the same 30 essays with the method on which they were trained on day two. Thus, all raters rated the same 30 essays using two methods, with essays randomly assigned within raters and across days. Raters used the data collection platform Qualtrics (2017) to access the ER-WR DRS questions for each essay and entered scores for the rubric essays in Google spreadsheets. Across both days, essays were randomized so that the order was not the same on day one as on day two and randomized across raters.

2.4 Many-facet Rasch Measurement

Linacre (1989) proposed a model-based approach for rater-mediated assessments. The MFRM model extends the traditional Rasch model (Rasch, 1980) to include facets, which are sources of variability thought to influence students' scores (Eckes, 2011), such as raters and tasks. The MFRM focuses on individual elements of each facet, allowing for direct feedback to individual raters concerning their performance and specific information concerning the difficulty and performance of the five ER-WR scoring elements. In addition, MFRM allows users to investigate specific rater behaviors such as central tendencies, halo, and randomness effects. Raters exhibit the central tendency effect when they overuse the middle category of a rating scale while avoiding the extreme categories (Myford &Wolfe, 2003), whereas, halo effect manifests when a rater fails to discriminate among conceptually distinct aspects of an examinee's behavior (Saal, Downey, & Lahey, 1980). Myford and Wolfe (2004) define randomness as "a rater's tendency to apply one or more trait scales in a manner inconsistent with the way in which the other raters apply the same scales" (p. 206).

To estimate the extent to which rater and rating method effects are present in student ER-WR essay scores, rater facets are included in the MFRM model, allowing for statistical tests and descriptive measures of the variability in rater effects (Myford & Wolfe, 2004). The MFRM applied in this study which include examinee, rater, and rubric element facets, is specified as:

$$ln\left[\frac{P_{nij(x=k)}}{P_{nij(x=k-1)}}\right] = \theta_n - \delta_i - \alpha_j - \tau_{ijk}, \quad (1)$$

where:
$P_{nij(x=k)}$ is the probability of examinee n ($n = 1, \dots, N$) being rated in category k on element i ($i = 1, \dots, I$) by rater j ($j = 1, \dots, J$); $P_{nij(x=k-1)}$ is the probability of examinee n being rated in category k-1 on element i by rater j; θ_n is the ethical reasoning ability of examinee n;
δ_i is the difficulty of ER-WR rubric element i; α_j is the severity of rater j in how they rate examinees; and τ_{ijk} is the difficulty of score level k compared to score level k-1, which is free to vary across element i and rater j (Eckes, 2011), such that each rater can have a different rating behaviour per element. All facets are reported in logit scale. The MFRM model was estimated using joint-maximum likelihood estimation via FACETS 3.80.4 (Linacre, 2018). We assessed how raters interact with elements across rating methods. Rater separation reliability and index statistics were reported. Rater separation reliability is the proportion of observed variation in rater severity measures that is not attributed to measurement error. The rater separation reliability is mathematically given as

$$R = \frac{SD_T^2}{SD_O^2} \tag{2}$$

where R is the rater separation reliability, SD_T^2 is the true variance of the rater severity measures, and SD_O^2 is the observed variance of the rater severity measures. The rater separation index captures the number of distinct groups that raters can be classified into in terms of their severity measures. The rater separation index is specified as

$$H = \frac{4SD_T^2 + RMSE}{3RMSE} \tag{3}$$

where H is the rater separation index, SD_T^2 is the true variance of the rater severity measures, and RMSE is the root mean-square measurement error of the rater severity measures. These separation statistics are extended to other facets in the model such as examinee and element facets. Interested readers are referred to Eckes (2011) for extensive discussion of these separation statistics and other rater fit statistics.

To explore additional rater behaviours, we also included experience (novice and expert raters) as a dummy-coded variable to assess how rater experience interacts with elements across rating methods. Novice raters were defined as those having no rating experience with 8EQ essays and expert raters were those with at least three semesters of 8EQ essay rating experience. Expert raters have only been exposed to the rubric-rating approach for ER-WR essay scoring prior to this study, so it is important to determine whether prior rubric use influences use of the DRS.

Because the MFRM is an extension of the Rasch model, it is crucial to assess whether the scales have ordered categories or thresholds. Disordered thresholds pose a problem with the operation of the ratings defined by the thresholds (Andrich, 2011) and is

an indication that the rating scales does not function as designed (Eckes, 2011). In the context of MFRM, the ordering of the rating scale used could be a source of measurement error as raters may differ from each other in their interpretation of the ordering of the scales (Eckes, 2011). FACETS outputs the Rasch-Andrich thresholds (i.e., the ordering of the category thresholds) for each rater-by-element combination across the rating methods. We documented our findings in the result section.

2.5 Equating scores on the DRS and rubric rating methods

To compare scores from separate scoring approaches, there are two methods: concurrent calibration and separate calibration with facet anchoring. Concurrent calibration would add a method facet to Equation (1) and both the DRS and rubric method scores calibrated together at one time. However, we apply an equating approach for use with MFRM (Lunz & Suanthong, 2011). The benefit of an equating method is that the new approach (i.e., the DRS) can be put on the same scale as the original approach (i.e., the ER-WR rubric). Because MFRM has been used with the rubric (e.g., Holzman, 2018), we chose an equating approach over concurrent calibration. In addition, separate calibration allowed us to assess the empirical ordering of the categories across rating methods, a benefit that concurrent calibration does not offer. MFRM equating analyses were guided by steps documented by Lunz and Suanthong (2011) to equate multi-facet performance assessments: (a) establish a benchmark scale (i.e. the ER-WR rubric); (b) develop the new scale (i.e., the DRS rating method); (c) anchor the DRS scale to the benchmark rubric scale; (d) evaluate the quality of the displacement and fit, and un-anchor elements and raters that do not meet the criteria; and (e) produce examinee results using the criterion standard established on the benchmark scale.
Specific to this analysis, we analyzed the assessments rated using the rubric rating method as the benchmark scale. The rater and element facets were centered to have a mean of 0.00 logits. Next, we analyzed the essays rated using the DRS rating method, anchoring all 12 raters and five elements to the benchmark scale. According to Lunz and Suanthong (2011), "if anchored facet elements displace by more than 0.5 logits or misfit, it may be necessary to unachor them and let them float for the current administration." Raters and elements that did not meet the displacement and fit criteria (i.e., mean-square *Outfit* greater than 1.50) were unanchored and data reanalyzed until the displacement and fit criteria were met. Displacement is a measure that shows how much each facet moved from the rubric benchmark scale to the DRS scale (Lunz & Suanthong, 2011). Raters or elements that did not meet the displacement and fit criteria would be considered inconsistent in severity or difficulty, respectively.

3. Results

3.1 Model Comparison

Before proceeding to equate the results across rating methods, we compared two nested models to the specification in Equation 1 to determine the most appropriate rating scale structures. The model in Equation 1, which includes τ_{ijk} and that we will refer to as Model 1, allows rating scale category structures to vary across elements and raters. A simpler model (Model 2, with τ_{ik}) allows for rating scale category structures to vary across elements only. The model chi-square statistic (χ^2), Schwarz's (1978) Bayesian Information Criteria (BIC) or Akaike Information Criteria (AIC; Akaike, 1973, 1974) are all indices for assessing relative model-data fit between competing models, with smaller values indicating better fit. The change in χ^2 between a complex and less complex model ($\Delta\chi^2$) indicates difference in fit between models. If this change is significant, as indicated by $p < .05$, the more complex model provides better fit than the less complex model. Results presented in Table 1 indicate that Model 1, the most complex, provides significantly better fit than the more restrictive models for both the rubric ($\Delta\chi^2 = 698.12, \Delta df = 280, p < .01$, AIC = 5021.76, BIC = 5049.24) and DRS ($\Delta\chi^2 = 710.41, \Delta df = 204, p < .01$, AIC = 4152.68, BIC = 4180.16). We championed Model 1 because of superior fit; the remainder of the study presents results on this model only. That is, the benchmark scale in the equating process allows rating scale category structures to vary across elements and raters.

Table 1.

Model comparison test results

	DRS		Rubric	
	Model 1	Model 2	Model 1	Model 2
AIC	4152.68	4863.12	5021.76	5719.92
BIC	4180.16	4890.60	5049.24	5747.40
χ^2	4142.88*	4853.29*	5011.86*	5709.98*
df	1526	1730	1440	1720

Note. *$p < 0.01$; DRS = diagnostic rating system; χ^2 = chi-square statistic; *df* = degrees of freedom; AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria

Table 2.

Rater Calibrations and Anchor across Methods

		Rubric				DRS *(First Cycle)*					DRS *(Second Cycle)*				
Rater	Exp	Score Mean	Meas-ure (logits)	Outfit MnSq	SR-ROR	Score Mean	Meas-ure (logits)	An-chor	Outfit MnSq	Dis-place	Meas-ure (logits)	An-chor	Outfit MnSq	Dis-place	SR-ROR
1	N	4.33	0.07	0.91	0.59	3.65	0.07	A	1.02	0.26	0.07	A	1.04	0.23	0.50
2	N	3.49	0.20	0.72	0.48	2.97	0.20	A	0.75	0.01	0.20	A	0.80	-0.03	0.41
4	E	5.05	-0.16	1.18	0.50	6.40	-0.16	A	**1.87**	-0.42	-0.68	U	2.80	0.00	0.36
5	E	2.97	0.18	**2.11**	0.48	2.61	0.51		0.82	0.01	0.50		0.86	0.00	0.42
6	E	4.25	0.02	1.23	0.52	3.92	0.02	A	0.87	0.15	0.02	A	0.91	0.11	0.50
7	E	5.17	-0.28	1.26	0.48	4.86	-0.28	A	**2.07**	0.14	-0.20	U	2.07	0.00	0.45
8	N	4.47	0.04	0.94	0.59	3.39	0.04	A	**1.61**	0.42	0.47	U	1.29	0.00	0.50
9	N	5.04	-0.08	1.34	0.44	4.16	-0.08	A	0.83	0.15	-0.08	A	0.85	0.11	0.55
10	E	4.15	0.04	0.71	0.59	4.28	0.04	A	0.77	0.05	0.04	A	0.80	0.01	0.56
11	E	4.84	-0.19	1.32	0.56	4.92	-0.19	A	1.1	-0.02	-0.19	A	1.19	-0.09	0.52
12	E	5.00	0.02	1.18	0.49	5.29	0.02	A	**1.87**	-0.26	-0.30	U	2.40	0.00	0.42
13	E	3.83	0.14	1.18	0.51	4.88	0.14	A	1.07	-0.33	0.14	A	1.20	-0.40	0.54

Note. there is no Rater 3 because of rater attrition in the study. Only raters with complete data (i.e., raters 1-2, 4-13) were included in the analysis; N = novice; E = expert; *A* = anchored; U = unanchored; MnSq = mean square; Displace = displacement; Exp = experience; DRS = diagnostic rating system; SR-ROR = single rater-rest of the rater

3.2 FACETS Analyses

3.2.1 Equating of DRS and Rubric Rating Methods.

Tables 2 and 3 display the results of equating rater severity and element difficulty, respectively. We unanchored raters and elements that displace by more than 0.50 logits or have *Outfit* mean squares outside the acceptable range (i.e., greater than 1.50). In Table 2, four raters who did not meet the displacement or fit criteria were unanchored. In Table 3, three elements were anchored after fit and displacement were evaluated.

Table 3.

Element Calibrations and Anchor across Methods

Element	Rubric					DRS						
	Score Mean	Std Dev	Measure (logits)	SE	Outfit MnSq	Score Mean	Std Dev	Measure (logit)	SE	Anchor	Outfit MnSq	Displace
E1	6.35	1.65	-0.39	0.05	2.05	6.50	2.41	-0.66	0.04		2.51	-0.01
E2	4.17	2.07	0.13	0.04	0.61	4.58	1.94	0.13	0.04	A	0.56	0.14
E3	3.70	2.00	0.19	0.04	0.65	3.58	2.12	0.19	0.04	A	0.67	-0.07
E4	3.99	2.14	0.02	0.04	0.95	3.54	2.39	0.02	0.05	A	1.01	-0.13
E5	3.69	2.08	0.05	0.04	1.62	3.19	2.60	0.04	0.07		1.35	-0.01

Note. DRS = diagnostic rating system; MnSq = mean square; Displace = displacement; SE = standard error; *A* = anchored; Std Dev = standard deviation

3.2.2 Wright Variable Map.

Figure 3 presents the Wright maps for the two rating methods. The map displays the location of facets elements on the ER-WR performance assessment latent scale. The first column in each display of the Wright map shows the logit measure, with a range of -2 to +2, for which rater severity and element difficulty are centered at zero. The examinee ability was not centered at zero. Rater severity is in the second column of the map. Raters that appear lower in the column are more lenient, while more severe raters appear at the top of the column. The third column displays the ability of the examinees. High-achieving examinees on the ER-WR performance assessment appear at the top of the column; conversely, low-achieving examinees appear at the bottom of the column. The elements are shown in the fourth column. More difficult elements

appear at the top of the column, whereas, easier elements appear at the bottom of the column.

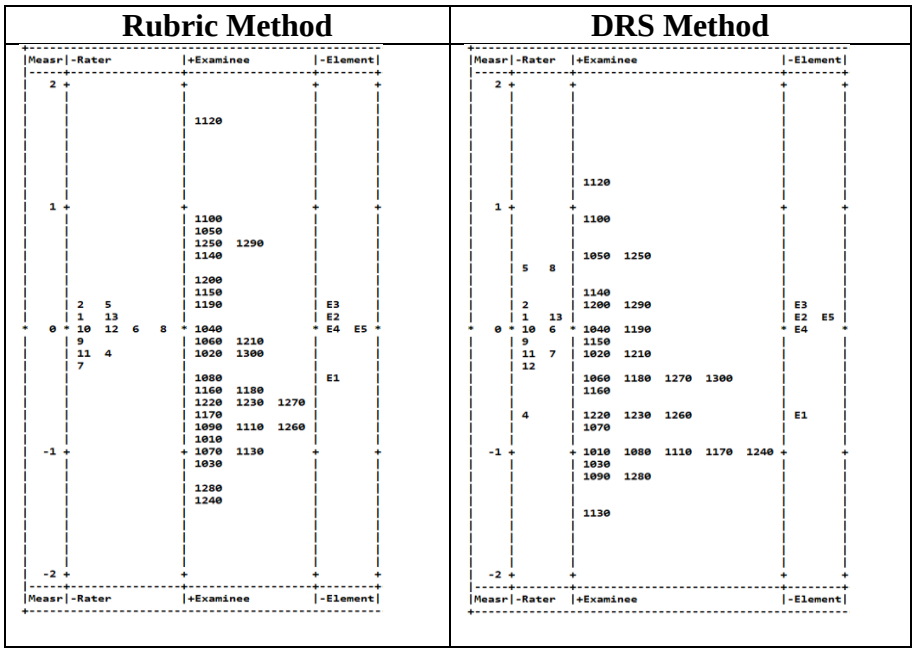

Figure 3:
Wright map for rater, examinee, and element facets across rating methods. The left panel shows the map for the rubric rating method. The right panel shows the map for the DRS rating method.

RQ1. Are rater effects mitigated with the use of the DRS method of scoring ER-WR essays?

After the equating process, we computed MFRM summary measures (on the logit scale), measures of fit, and separation statistics of raters across rubric and DRS rating methods, which are presented in Table 4. One of the goals of employing the ER-WR DRS was to reduce raters' subjective interpretation of score criteria when using the ER-WR Rubric. The fixed-effect chi-square of the rater facet evaluates the null hypothesis that, after controlling for measurement error, there are no differences in rater severity of ER-WR ratings. A statistically significant chi-square (i.e., $p<.05$) suggests

that at least two raters are significantly different in their leniency/severity (Myford & Wolfe, 2004). The fixed-effect chi-square statistic confirms a statistically significant difference in terms of severity/leniency when using the ER-WR Rubric ($\chi^2 = 55.1, df = 11, p < .01$) and ER-WR DRS ($\chi^2 = 217.7, df = 11, p < .01$) methods. The current model used in this study permits the investigation of whether raters overused of the middle category. A nonsignificant fixed-effect chi-square statistic is an indication of a group-level central tendency effect. The fixed-effect chi square statistics indicate that the raters did not exhibit any group level central tendency effects when using the ER-WR Rubric and ER-WR DRS. The high examinee separation statistics reported in Table 4 also confirms that the raters did not exhibit group-level central tendency across the two rating methods ($R_{DRS} = 0.97$ and $R_{Rubric} = 0.98$).

The rater separation reliability reported in Table 4 reflects the unwanted variation between raters in their levels of severity/leniency (Myford & Wolfe, 2003), with rater separation reliability close to zero representing the ideal scenario. Judging by the reliability of separation statistic, raters were more homogeneous in their ratings when using the rubric (R = 0.79) compared to the DRS method (R = 0.95). The separation index suggests the presence of at least two distinct groups of raters when using the ER-WR Rubric (H = 2.89) and at least five distinct groups of raters when using the ER-WR DRS (H = 5.97).

The level of rater consistency in the use of the different rating methods was assessed using the *Infit* and *Outfit* statistics. Myford and Wolfe (2003) define fit indices as the degree to which observed ratings match the expected ratings that are generated by the MFRM. In the present study, the average mean-square statistics for both the ER-WR DRS (*Infit* = 1.04 and *Outfit* = 1.35) and ER-WR Rubric (*Infit* = 1.04 and *Outfit* = 1.17) show more variation in ratings than expected. It should be noted that *Outfit* statistic is an unweighted mean-square residual that is sensitive to unexpected outlier ratings, whereas, *Infit* statistic is a weighted mean-square residual that is less sensitive to unexpected outlier ratings. The average *Infit* statistic was similar for both rating methods but the *Outfit* statistic slightly differs.

Table 4.
Overall Measurement Report by Facets

	Rubric Method			DRS Method		
	Rater	Examinee	Element	Rater	Examinee	Element
Measure (logit)						
M	0.00	-0.25	0.00	0.00	-0.39	0.00
SD	0.15	0.75	0.23	0.33	0.67	0.38
N	12	30	5	12	30	5
Infit Mean-Square						
M	1.04	1.10	1.12	1.04	1.07	1.12
SD	0.23	0.49	0.50	0.20	0.40	0.48
Outfit Mean-Square						
M	1.17	1.17	1.17	1.35	1.35	1.35
SD	0.37	0.62	0.63	0.69	1.21	1.10
Separation Index, H	2.89	8.97	6.95	5.97	7.70	10.6
Separation Reliability, R	0.79	0.98	0.96	0.95	0.97	0.98
Fixed Effect Chi-square	55.1*	1170.1*	80.5*	217.7*	837.6*	289.4*
Degrees of freedom	11	29	4	11	29	4

Note. *$p<.01$; *M* = mean; *SD* = standard deviation; *N* = number of raters/examinees/elements; DRS = diagnostic rating system

After establishing that raters exhibited differential rater leniency/severity in the two rating methods, we assessed the individual rating behaviors of each rater when using the ER-WR Rubric compared to the ER-WR DRS method. First, we evaluated whether raters applied the rating scales in a similar fashion to the way other raters applied the same rating scales. The single rater-rest of the rater (SR-ROR) correlation is one measure of inter-rater reliability, the MFRM version of the point-biserial correlation (Linacre, 2003). SR-ROR reflects the degree to which a rater's ratings (i.e., the single rater) are consistent with the ratings of the rest of the raters (i.e., the rest of the raters). Myford and Wolfe (2003) indicate that low SR-ROR correlations are those less than 0.30 and high SR-ROR correlations are those greater than 0.70. The SR-ROR correlations are presented in Table 2. The findings show that all raters across both methods had SR-ROR correlations between 0.36 and 0.59, indicating adequate

inter-rater reliability using the ER-WR Rubric and DRS methods. Because SR-ROR correlations of these raters are comparable to other raters, we can conclude that these raters did not exhibit high degree of randomness.

Finally, we evaluated whether raters display halo effects. The findings as presented in Table 2 show that some raters exhibited large differences in rating behaviors across rating methods. Raters' *Infit* and *Outfit* fit statistics provide information on rater consistency in the use of the rating scales. Large *Infit* and *Outfit* statistics are an indication of inconsistency or other rater biases in the use of the rating scales. The results indicate that, when compared to other raters, Rater 5 (*Outfit* = 2.11) display the most inconsistency in ratings on the ER-WR Rubric while Rater 4 (*Outfit* = 2.80), Rater 7 (*Outfit* = 2.07), and Rater 12 (*Outfit* = 2.40) display large differences between their observed and expected ratings on the ER-WR DRS. Myford and Wolfe (2004) suggest that when element difficulties are allowed to vary, the ratings of raters who exhibit halo effects will be very different from the expected ratings and will result in *Infit* and *Outfit* mean-square statistics that are significantly greater than one. Our result indicates that Rater 5 and Raters 4, 7, and 12 may be exhibiting halo effect when using the ER-WR Rubric and ER-WR DRS, respectively. Linacre (2002) cautions that fit indices greater than 2.0 could "distort or degrade the measurement system." These four raters fall within the range of questionable ratings judging by their *Outfit* statistic values. Noteworthy, although not reported in Table 2, the *Infit* statistic of all 12 raters falls within the range of acceptable fit (i.e. between 0.50 and 1.50) in both rating methods.

RQ2. How does examinees' proficiency compare across rating methods?

Despite the differences in rater severity, the findings show that examinee proficiency levels are comparable between rating methods. The FACETS analysis revealed that examinees separation index was larger for ER-WR Rubric (H = 8.97) than for ER-WR DRS (H = 7.70). This suggests that examinees can be classified into nine distinct classes for rubric, but into almost eight classes for DRS. As reported in Table 4, the fixed-effect chi-square statistics for both rating methods reveal that after controlling for measurement error, all examinees did not have the same level of performance. Similarly, the reliability of separation statistics was close to 1 for both rating methods. This means that the assessment essay was able to differentiate examinees in terms of their performance regardless of the rating method applied.

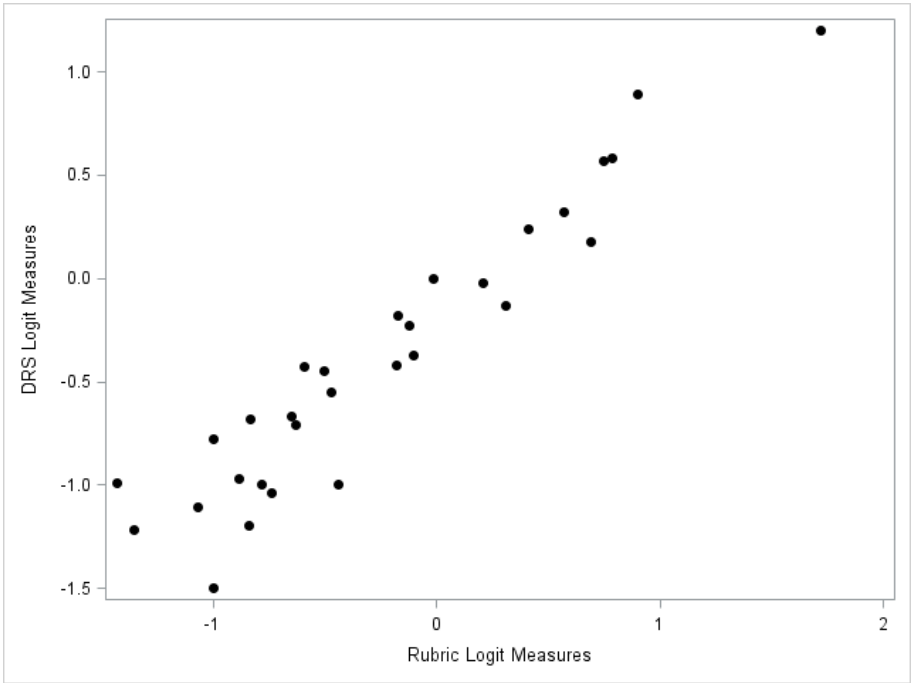

Figure 4.
Scatterplot of examinee logit measures across methods ($R^2 = 0.90$), with rubric logit measures on the horizontal axis and DRS logit measures on the vertical axis.

Next, we compared examinee proficiency (i.e., on the logit scale) based on ratings received on the rubric and DRS rating methods. As presented in Figure 4, there is a positive linear relationship between the performance of examinees when rated on the rubric and DRS with a high correlation coefficient ($r = 0.95$, $R^2 = 0.90$). This implies that the ordering and measures (logits) of examinee achievement were similar on both rating methods.

RQ3. Does rater experience impact the quality of ratings across rating methods?

This study used a mixture of expert and novice raters. We ran a separate analysis which included experience as a facet. The inclusion of the experience facet in our analysis allowed for the investigation of how rater experience interacts with ratings awarded. The experience facet served as a dummy variable. Novice and expert raters were trained together and were given the same tasks. The experience facet report is presented in Table 5. When using the ER-WR Rubric, the summary statistics of scores awarded reveal that expert and novice raters, on average, awarded similar ratings with

an observed average rating of 4.41 (SD = 1.60) and 4.33 (SD = 1.64), respectively. However, a much higher differential mean ratings were observed when using the ER-WR DRS. Under this method, novice raters awarded lower ratings with a mean rating of 3.54 (SD = 1.35) compared to 4.65 (SD = 1.81) for expert raters. We note that the average ratings awarded by expert raters were similar across rating methods. Also, the standard deviation of the ratings indicates that expert raters used a wider range of the rating scale. Comparison across the two rating methods show that novice raters awarded lower scores when using the ER-WR DRS compared to ratings when using the ER-WR Rubric. In contrast, the ratings awarded by expert raters was higher on the ER-WR DRS.

Table 5.
Experience facet measurement report

Exp	Rubric					DRS				
	Score Mean	Measure (logit)	SE	Infit MnSq	Out-fit MnSq	Score Mean	Meas-ure (logit)	SE	Infit MnSq	Outfit MnSq
Novice	4.33	0.00	0.03	0.96	0.98	3.54	0.00	0.04	0.90	1.00
Expert	4.41	0.00	0.02	1.11	1.27	4.65	0.00	0.03	1.11	1.53

Novice and expert raters *Infit* and *Outfit* statistics were within acceptable range on the ER-WR Rubric, but the *Outfit* statistic of expert raters was outside the acceptable range *(Outfit* = 1.53) on the ER-WR DRS. A graphical illustration of the raters' measure scores presented in Figure 5 show four raters (Raters 4, 5, 8, and 12) who are more or less severe/lenient across rating methods. We note that three out of these four raters were expert raters. These findings might suggest that expert raters appear to make some other judgments not captured when using the ER-WR DRS rating method.

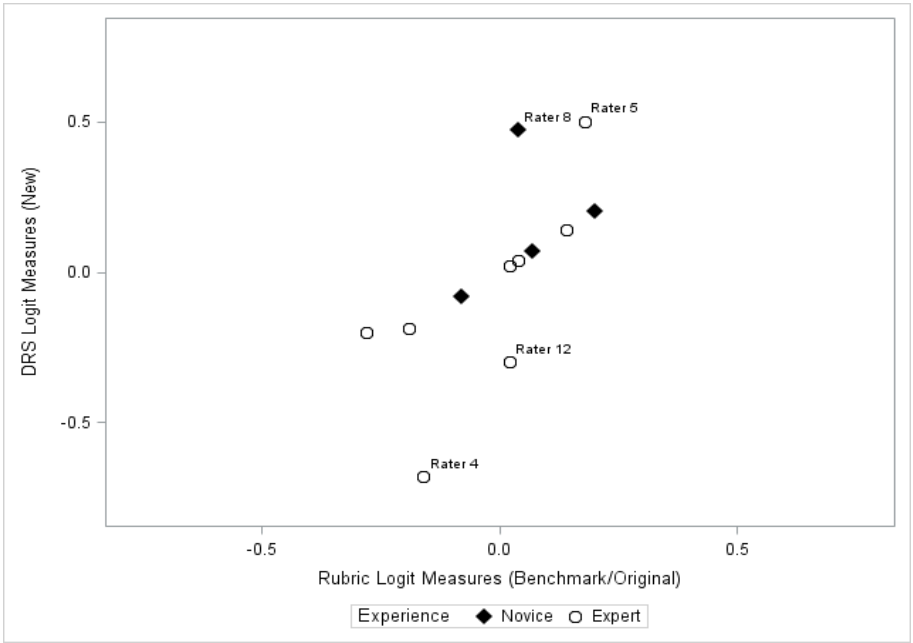

Figure 5.
Scatterplot of rater logit measures across rating methods and experience, with novice raters indicated using solid diamonds and open circles representing expert raters.

The MFRM allows for the investigation of systematic patterns between facets in order to gain additional information about the behavior of one facet when crossed with other facets. Table 6 presents the results of the bias/interaction analyses. Bias estimates of zero indicate that based on the model, the observed scores and expected scores are equal. Positive bias estimates indicate that the observed scores are greater than the expected scores based on the model, whereas, negative bias estimates indicate that the observed scores are lower than the expected scores. A significant bias estimate was assessed using the *t*-statistic. Engelhard (2002) cautions that *t*-statistic greater than an absolute value of 2 is an indication of significant bias. A statistically significant *t*-statistic means that the combination of facets led to an unexpectedly high or low rating (Myford & Wolfe, 2003).

An element-by-experience analysis was conducted to gain more insight on the potential sources of bias. The findings presented in Table 6 show that novice and expert raters, on average, maintained uniform level of severity/leniency across the five levels on the ER-WR Rubric ($\chi^2 = 3.3,\ df = 10, p$ =0.9735). This result agrees with Figure 6 which shows that none of the expert and novice raters were flagged for significant

severity or leniency bias for ratings awarded using the ER-WR Rubric. However, the study found a statistically significant interaction between element and experience on the ER-WR DRS ($\chi^2 = 41.3,\ df = 10,\ p$ <.01). When using the DRS, Element 1 rated by novice raters was flagged for significant severity bias (*Bias* = -0.24, *t* (119) = -2.90, *p* = 0.0044) while ratings awarded by expert raters were flagged for significant severity bias on Element 2 (*Bias* = -0.17, *t* (239) = -3.17, *p* = 0.0044) and significant leniency bias on Element 4 (*Bias* = 0.19, *t* (239) = -3.17, *p* = 0.0044). This tells us that, on average, novice raters may be interpreting Element 1 differently while, as a group, expert raters may have different interpretations to Elements 2 and 4.

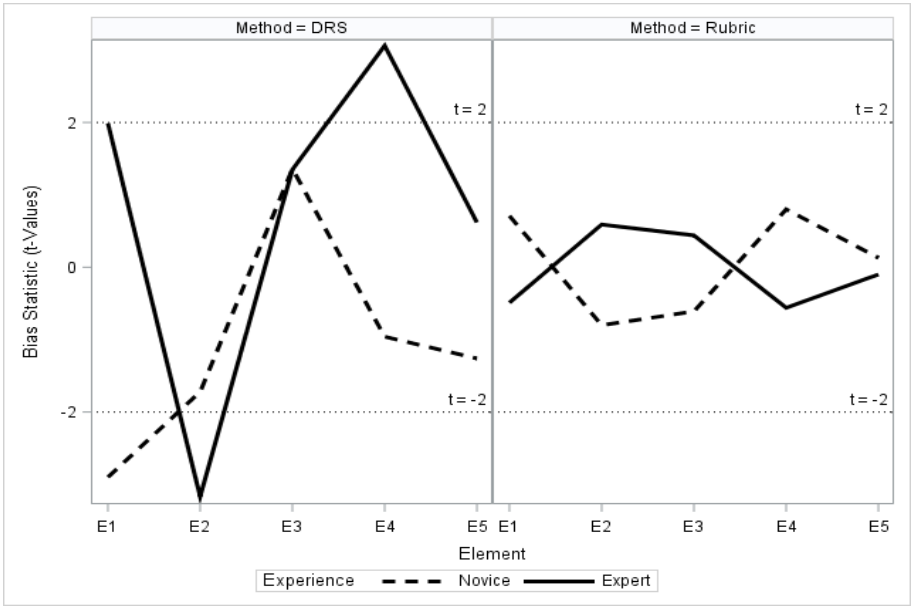

Figure 6.
Bias/Interaction Plot between Rater Experience and Element by Rating Method

RQ4. How did elements on the rating scale function across rating methods?
It was important to evaluate the difficulty of the elements and how raters applied these elements across rating methods. As indicated by the fixed-effect chi-square statistics of the element facet in Table 4, the findings suggest that the difficulty levels of at least two elements in this assessment differed significantly for both the ER-WR Rubric ($\chi^2 = 80.5,\ df = 4$, p < .01) and ER-WR DRS ($\chi^2 = 289.4,\ df = 4$, p < .01). We observed that the separation index is higher for the ER-WR DRS (H = 10.6) than for

the ER-WR Rubric (H = 6.95). This means that the ER-WR DRS elements could be classified into about ten-and-half distinct classes. Both separation indexes are higher than the number of elements on the ER-WR performance instrument. Eckes (2011) notes that a separation index greater than the number of elements could be due to large "true" standard deviations of the elements or due to a large number of observations. As reported in Table 3, the standard deviations of these elements are around 2 points from the observed ratings, on average. This could justify adding more elements to narrow the spread (Eckes, 2011).

After establishing that elements differ significantly, we compared the measurement results within and between the ER-WR Rubric and ER-WR DRS rating methods. The results in Table 3 (also displayed in Figure 3) show that Element 3 is the most difficult element on the ER-WR performance instrument for both rating methods. We also observe that Element 1 is the easiest on the assessment under both rating methods. Although raters agree that Element 1 was the easiest element, the Outfit statistic suggests that raters were more inconsistent when rating examinees on Element 1 ($Outfit_{RUBRIC}$ = 2.05 and $Outfit_{DRS}$ = 2.51). The Outfit statistic of Element 5 rated on the ER-WR Rubric also falls outside the acceptable range.

Table 6.
Summary Statistics for the interaction analysis

	Rater-by-Element Interaction		Element-by-Experience Interaction	
	Rubric	DRS	Rubric	DRS
N combinations	60	60	10	10
M Bias	-0.01	-0.04	0.00	-0.04
SD Bias	0.25	0.42	0.04	0.15
% large *t*-statistics[a]	20.00	26.67	0.00	30.00
Fixed Effect Chi-square	129.8*	171.6*	3.3	41.3*
Degrees of freedom	60	60	10	10

Note. *$p<.01$; M = mean; SD = standard deviation; DRS = diagnostic rating system; [a]Percentage of absolute *t*-statistics greater than or equal to 2.

In our study, the rater-by-element bias analysis assessed whether raters exhibited a similar level of severity or leniency when rating elements using the ER-WR DRS and ER-WR Rubric rating methods. We hypothesized that raters' bias will be less severe/lenient on the ER-WR DRS because it requires less cognitive load on raters and is more streamlined. The results indicate a statistically significant rater-by-element interaction on the ER-WR DRS ($\chi^2 = 171.6,\ df = 60, p < .01$) and on the ER-WR

Rubric ($\chi^2 = 129.8,\ df = 60,\ p < .01$). This implies that raters did not maintain a uniform level of severity/leniency across the rating elements. The rater-by-element interaction is illustrated in the plot in Figure 7. The vertical axis is the raters' test statistic of the bias severity/leniency across the elements. In this figure, it is seen that raters exhibited differential levels of bias within elements and between rating methods. For example, on Element 1 (top panel of Figure 7), four raters (Raters 1, 12, 5, and 8) displayed differential leniency/severity on the ER-WR DRS with statistically significant bias ($|t| \geq 2$), denoted by faint dashed lines in Figure 7. We further illustrate rater differential bias on different elements using Rater 1. It can be seen that Rater 1 did not display significant bias when rating Element 3 using the ER-WR Rubric ($Bias = 0.10, t(29) = 0.68, p = 0.5019$) and ER-WR DRS ($Bias = 0.07, t(29) = 0.53, p = 0.6001$). However, Rater 1 displayed dissimilar rating patterns on Element 5 across the rating methods as shown in the bottom panel of Figure 7. Rater 1 exhibited significant leniency bias when rating with the ER-WR Rubric ($Bias = 0.44, t(29) = 2.41, p = 0.0255$) but awarded severe ratings on the ER-WR DRS ($Bias = -0.29, t(29) = -2.84, p = 0.0082$). Similar results were found when examining the category characteristic curves (i.e., threshold ordering), so we chose to provide the bias results for brevity.

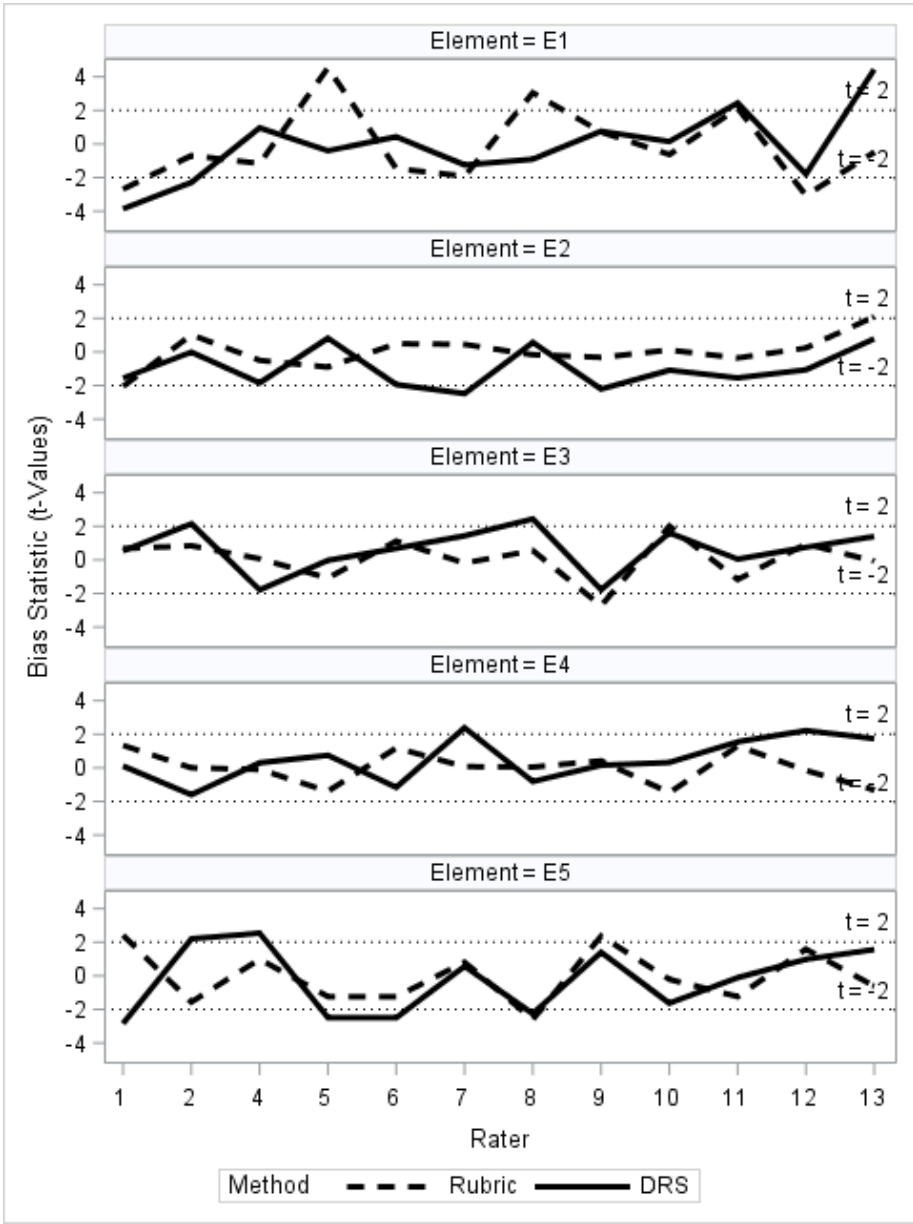

Figure 7.
Bias/Interaction Plot between Rater and Rating Method by Rating Element

4 Discussion

The validity of the inferences made from performance assessments depends on the quality of the ratings. This study was established to illustrate the DRS concept, compare examinee proficiency on the DRS and the traditional rubric rating methods, as well as assess how raters fare on both rating methods. Fully-crossed designs are often desired in performance assessments because more measurement information of each examinee is obtained. One of the strengths of this study is the implementation of a fully-crossed rating design of 12 raters and 30 examinees resulting in 12 ratings per examinee on each element. Overall, the DRS and Rubric rating methods were good at differentiating the ethical reasoning ability of examinees. The correlation between examinee proficiency on the DRS and rubric rating methods was strong. The high correlation in examinee performance level lends us to believe that the DRS functions in a similar fashion as the traditional rubric in estimating examinee abilities. In addition, the examinee separation statistics were similar across the rating methods. The practical implication of the examinee performance ordering and separation statistics is that raters are believed to assign valid ratings on the traditional rubric, which is the benchmark scale in which comparisons are made from using the DRS.

Consistency of ratings (and raters) is important in any performance assessment. The DRS was designed to reduce rater subjectivity, provide a more streamlined approach to rating with less rater cognitive load, and provide validity evidence that the DRS approach can be used in lieu of the traditional rubric-based scoring approach. Rater training was an important consideration in this study in order to ensure that raters have a shared understanding of the ER-WR Rubric and ER-WR DRS rating methods. Despite training, the findings from the current study show evidence of variation of raters in scoring propensity as evidenced in the differences in rater severity/leniency on the two rating methods we employed – a finding which correlates with the inconsistency in their ratings. Our findings on rater severity are consistent several studies (e.g., Weigle, 1998; Eckes, 2020) which found that rater training did not eliminate the variation in rater severities. Although none of the raters exhibited any effects due to group-level central tendency or randomness, about one-quarter of the raters display high inconsistency in their ratings on the DRS. This could be as a result of halo effect which occurs when ratings awarded by raters are different from the expected ratings (Myford & Wolfe, 2004) or because some expert raters had spent at least three rating sessions using the rubric. The fit statistics suggest that these raters could be interpreting the elements differently by either awarding lower (or higher) than expected ratings. It is important to note, as highlighted by Marais and Andrich (2011), that halo effect is not only a property of individual raters but could be due to the common structure of the rating scales. Marais and Andrich (2011) discuss an approach to detect halo effect common to all raters.

Across rating methods, raters were more homogenous when using the traditional rubric compared to the DRS. This was also evidenced by the rater separation statistics as the number of distinct classes of the raters under the DRS was twice the number of distinct classes under the traditional rubric method. A closer look at the characteristics

of the raters reveal some systematic patterns when rater experience was considered. On average, expert raters are likely to award similar ratings across rating methods but tend to use a wider range of rating criteria compared to novice raters. This finding parallels Cumming (1990) and Wolfe, Kao, and Ranney (1998). Most importantly, there were no significant bias when rater experience was crossed with rating elements on the ER-WR Rubric rating method but three raters (one novice and two expert raters) were flagged for significant bias on ER-WR DRS. When using traditional rubrics, previous studies (e.g., Kaufman, Baer, Cole, & Sexton, 2002) found that novice raters were inconsistent in ratings when compared to expert raters. The DRS proves otherwise as we found that novice raters display high consistency on the ER-WR DRS. This may not be unsurprising because the DRS is designed to guide raters in a streamlined manner, thus will be especially beneficial to less experienced raters. This has implications for practitioners who employ performance assessments. While the study shows that examinee proficiency estimates can be accurately estimated using the DRS, it also reveals that some expert raters were inconsistent when using the DRS. However, in the face of difficulties in recruiting experienced raters, the DRS approach will prove invaluable when less experienced raters are recruited. Practically, this means that the DRS appears to benefit inexperienced raters who may need a more streamlined process of rating.

One possible explanation why the two expert raters in this study show significant bias when using the DRS could be because they apply some prior knowledge or other criteria not captured by the system. Perhaps this might be one of the shortcomings of the DRS. It is possible that the current DRS lacked comprehensive explicit questions that these raters would like to judge the quality of the ER-WR essays on. The lack of explicitness of criteria may increase the bias in ratings. Also, it may hinder raters from providing the examinees with quality feedback. To this end, careful thoughts must be taken at the design stage of the DRS. The explicit rating criteria must be comprehensive. In addition, potential factors that may affect the rating process needs to be considered at the design stage as well. Bejar (2012) called for a strong consideration of rater cognition at the design stage of rubric development. These considerations should be extended to the DRS. We believe that the DRS will benefit raters if their inputs are considered when designing the system. Pilot testing the system with a few essays and eliciting feedback from raters will prove worthwhile in ensuring that all possible explicit questions are captured by the system. The piloting phase could also explore the rating behaviors and beliefs of raters in interpreting rating scales especially between novice and expert raters. This additional information will benefit practitioners especially in selecting suitable raters.

Limitations and Future Research

There are a few limitations in this study. First, we employed a fully-crossed design in the current study. Fully-crossed designs may not be practically possible because of the cost implication. Also, in fully-crossed designs raters may experience effects such

as rater fatigue if the sample size is large. It is unknown if the findings here can be generalized to incomplete rating designs. It may be important to assess the performance of the DRS with an incomplete rating design. Second, the current study did not capture the time it took raters to rate the 30 essays on the DRS and rubric rating methods. The time it took to rate essays would be valuable information for buying into the DRS, as it may be more cost efficient to use a system that reduces rating time and requires less rater cognition.

Although the order of rating using the DRS or rubric method was randomly assigned, raters only had a day between the two rating methods. Potentially, this one-day "washout" between using the rating methods may have been too short. This might explain some of the differences we found across rating methods. Future studies may want to have a longer washout period and use different artifacts between rating methods. Because raters saw the same essays on both rating days, they may not have read the essay as closely on the second day, perhaps distorting quality of day-two's ratings. However, the DRS is meant to serve as an alternative to ER-WR rubrics and real-world use of the DRS may not include both methods. Because the purpose of this study was to compare the two approaches, we made raters use both.

Another direction for future research is the comparison between the DRS and comparative judgment methods. Compared to traditional rubrics, Steedle and Ferrara (2016) suggest that comparative judgment approach could reduce rater fatigue due to decrease in cognitive load associated with the relative judgment of essays. It would be interesting to investigate how the DRS can be applied alongside the comparative judgment approach. Finally, the scope of this study did not include the cost-benefit analysis of employing the DRS compared to using traditional rubric. We hope to examine this in the future. Despite these limitations, the DRS appears to offer a useful option for rater-mediated assessments

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory, 1973* (pp. 268-281). Publishing House of the Hungarian Academy of Sciences.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716-723.

Ames, A. J., & Luecht, R. (2018). Item development. In B. B. Frey (Ed.) *The SAGE Encyclopedia of Educational Research, Measurement and Evaluation* (pp. 894-898). Thousand Oaks, CA: SAGE.

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert review of pharmacoeconomics & outcomes research*, *11*(5), 571-585.

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning and Assessment*, *10*(3).

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2-9.

Cohen, Y., Levi, E., & Ben-Simon, A. (2018). Validating human and automated scoring of essays against "True" scores. *Applied Measurement in Education*, *31*(3), 241-250.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31-51.

Curtis, N., & Ames, A. (2018). *Evaluating the Utility of the Diagnostic Rating System for Performance Assessment*. Annual Conference of the National Council on Measurement in Education (NCME), Paper presentation; April 13th-16th, 2018. New York, NY.

Downing, S. M. (2006). Twelve steps for effective test development. In *Handbook of test development*, *3*, 25.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*, 88-115.

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

Eckes, T. (2020). Rater-Mediated Listening Assessment: A Facets Modeling Approach to the Analysis of Raters' Severity and Accuracy When Scoring Responses to Short-Answer Questions. *Psychological Test and Assessment Modeling*, *65*(4), 449-471.

Engelhard Jr, 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.

Engelhard, G. (2002). Monitoring raters in performance assessments. *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*, 261-287.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing*, *27*(3), 317-334.

Ethical Reasoning in Action (n.d.). Eight key questions. Retrieved from https://www.jmu.edu/ethicalreasoning/8-key-questions.shtml

Holzman, Madison A., "Evaluating rater effects in the context of ethical reasoning essay assessment: An application of the many-facets Rasch measurement model" (2018). Dissertations. 192. https://commons.lib.jmu.edu/diss201019/192

James Madison University (2014). James Madison University's ethical reasoning rubric. Ethical Reasoning in Action. https://www.jmu.edu/ethicalreasoning/Docs/Ethical%20Reasoning%20Rubric%20-%20JMU%20-%20Final.pdf

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, *20*(2), 171-178.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*, 81-96.

Leighton, J. P. (2019). The risk–return trade-off: Performance assessments and cognitive validation of inferences. *British Journal of Educational Psychology*, *89*, 441–455.

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

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

Linacre, J.M. (2003). Computing the "Single Rater-Rest of Raters" (SR/ROR) Correlations. Appendix A in C. Myford & E. Wolfe: Detecting and Measuring Rater Effects, *Journal of Applied Measurement, 4,* 421-2.

Linacre, J. M. (2018) Facets computer program for many-facet Rasch measurement, version 3.80.4. Beaverton, Oregon: Winsteps.com

Linder, F., Ames, A., Hawk, W., Pyle, L., Fulcher, K., & Early, C. (2020). Teaching ethical reasoning: Program design and initial outcomes of a university-wide ethical reasoning program. Teaching Ethics .19(2):147-169. https://doi.org/10.5840/tej202081174

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational researcher*, *20*(8), 15-21.

Lunz, M., & Suanthong, S. (2011). Equating of multi-facet tests across administrations. *Journal of Applied Measurement*, *12*(2), 124-134.

Marais, I., & Andrich, D. (2011). Diagnosing a common rater halo effect using the polytomous Rasch model. *Journal of Applied Measurement*, *12*(3), 194-211.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189-227.

Pollitt, A. (2004). *Let's stop marking exams.* Paper presented at the IAEA Conference, Philadelphia, PA. Retrieved from https://www.cambridgeassessment.org.uk/Images/109719-lets-stop-marking-exams.pdf

Popham, W. J. (1990). *Modern educational measurement: A practitioner's perspective*. IOX Assessment Associates.

Qualtrics (2017). Available from https://www.qualtrics.com. Provo, UT.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

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.

Sanchez, E.R.H., Fulcher, K.H., Smith, K., Ames, A., & Hawk, W.J. (2017). Defining, teaching, and assessing ethical reasoning in action. *Change: The Magazine of Higher Learning*, *49*(2), 30-36.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of statistics*, *6*(2), 461-464.

Sims, M. E., Cox, T. L., Eckstein, G. T., Hartshorn, K. J., Wilcox, M. P., & Hart, J. M. (2020). Rubric Rating with MFRM versus Randomly Distributed Comparative Judgment: A Comparison of Two Approaches to Second-Language Writing Assessment. *Educational Measurement: Issues and Practice*, *39*(4), 30-40.

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, *29*(3), 211-223.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, *34*(4), 273.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language testing*, *15*(2), 263-287.

Weigle, S. C. (2011). Validation of automated scores of TOEFL iBT® tasks against nontest indicators of writing ability. *ETS Research Report Series*, *2011*(2), i-63.

Welch, C. (2006). Item and prompt development in performance testing. *Handbook of test development*, 303-327.

Wind, S. A., Wolfe, E. W., Engelhard Jr, G., Foltz, P., & Rosenstein, M. (2018). The influence of rater effects in training sets on the psychometric quality of automated scoring for writing assessments. *International Journal of Testing*, *18*(1), 27-49.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*(1), 83-106.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465-492.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *R & D Connections*, *21*(2), 1-11.