2018-04-01

Rubric Rating with MFRM vs. Randomly Distributed Comparative Judgment: A Comparison of Two Approaches to Second-Language Writing Assessment

Maureen Estelle Sims
Brigham Young University

Rubric Rating with MFRM vs. Randomly Distributed Comparative Judgment: A Comparison of Two Approaches to Second-Language Writing Assessment

Maureen Estelle Sims

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Troy L. Cox, Chair
Grant Eckstein
K. James Hartshorn

Department of Linguistics and English Language

Brigham Young University

ABSTRACT

Rubric Rating with MFRM vs. Randomly Distributed Comparative Judgment:
A Comparison of Two Approaches to Second-Language
Writing Assessment

Maureen Estelle Sims
Department of Linguistics and English Language, BYU
Master of Arts

The purpose of this study is to explore a potentially more practical approach to direct writing assessment using computer algorithms. Traditional rubric rating (RR) is a common yet highly resource-intensive evaluation practice when performed reliably. This study compared the traditional rubric model of ESL writing assessment and many-facet Rasch modeling (MFRM) to comparative judgment (CJ), the new approach, which shows promising results in terms of reliability and validity. We employed two groups of raters—novice and experienced—and used essays that had been previously double-rated, analyzed with MFRM, and selected with fit statistics. We compared the results of the novice and experienced groups against the initial ratings using raw scores, MFRM, and a modern form of CJ—randomly distributed comparative judgment (RDCJ). Results showed that the CJ approach, though not appropriate for all contexts, can be valid and as reliable as RR while requiring less time to generate procedures, train and norm raters, and rate the essays. Additionally, the CJ approach is more easily transferable to novel assessment tasks while still providing context-specific scores. Results from this study will not only inform future studies but can help guide ESL programs to determine which rating model best suits their specific needs.

Keywords: rubric rating, many-facet Rasch measurement model (MFRM), comparative judgment (CJ), reliability of ESL writing assessment, practicality of ESL writing assessment

ACKNOWLEDGMENTS

It is an honor to acknowledge those who have helped complete this work. To Matthew Wilcox and Judson Hart, many thanks for the hours spent helping to set up and analyze the data for this study. I am deeply grateful to Rachel Armstrong, for her many hours of diligent *i*-dotting and *t*-crossing. I am indebted to those anonymous editors who reviewed the work and to the participants of the study. And finally, because this research was internally funded, I want to express special appreciation to the English Language Center at Brigham Young University for their cooperation.

TABLE OF CONTENTS

1. Introduction

One aim of writing assessment is to develop valid, reliable, and practical means of directly evaluating student writing, especially in large-scale assessment situations such as placement or proficiency testing. Assessment cannot be valid without reliability, yet finding balance between reliability and practicality is notoriously difficult. While holistic rating enjoys some popularity as a practical approach to essay assessment, this practicality can come at the expense of reliability. Innovations in analytic and trait-based rubric scoring have led to increased reliability but often at the expense of practicality. Some may argue that a modern approach to direct writing assessment, which predates analytic rubric rating, has the potential to allow for both practicality and reliability.

This approach, called comparative judgment (CJ), provides a method of direct writing assessment through paired comparisons in which raters are simply asked to select the better of two essays. Rater selections are aggregated and result in a rank-ordered scale, which demonstrates the relative distance between each of the essays. In essence, CJ presents a potentially measurable approach to direct writing assessment (Whitehouse & Pollitt, 2012).

Writing assessors may be rightly skeptical of CJ because it promises efficient, reliable results, essentially norm-referencing writing assessment, and some evidence suggests some versions of it are seriously flawed (Bramley & Wheadon, 2015); however, others view it as a promising alternative to traditional rubric rating (also known as RR; Pollitt, 2004; Steedle & Ferrara, 2016). Given the renewed interest in CJ and the need for more practical, yet reliable, approaches to direct writing assessment, additional research is needed to determine its utility. This research compares CJ with the traditional analytic RR system and examines indices of both reliability and practicality to determine whether CJ may be a viable alternative to RR.

2. Literature Review

It is well established in the writing community that productive tasks (e.g., essays) are a preferred method to validly assess writing skill over discrete response tasks (e.g., multiple choice; Greenberg, 1992; Huot, 1990; Yancey, 1999) and that a valid assessment is one in which inferences made and actions taken are adequately supported by both theoretical and empirical evidence (Messick, 1989). These principles are especially important as they pertain to writing placement or proficiency testing. Such direct assessment approaches, however, are notoriously difficult to rate reliably; history has shown that "as far back as 1880 it was recognized that the essay examination was beset with the curse of unreliability" (Breland, 1983, p. 1). Unreliable ratings indicate that raters are judging essays inconsistently—either at the individual level or in comparison to other raters—undermining assessment credibility, generalizability, and validity. Writing assessors value reliability because increased reliability reduces measurement error (Wiliam, 2001), which is an essential consideration in high-stakes testing, where test results have substantial consequences and need to be valid (O'Neill, 2011; Wiliam, 2001).

2.1 Increasing Reliability

In order to produce reliable essay scores, writing assessors have implemented numerous measures including creating rubrics, training raters on the rubrics, and using equalizing software to compensate for rater error and bias. Rubrics provide descriptions of expectations of quality, and raters assign examinees to locations on the scale to represent their relative ranking in relation to the traits being assessed (Myford & Wolfe, 2003). Holistic scoring rubrics include rating scale levels with descriptors for each level (Huot, O'Neill, & Moore, 2010). A more reliable version, analytic rubrics goes one step further and includes descriptors for varying criteria which can be measured individually and aggregated to reach a final score (Barkaoui, 2011).

Many claim a rubric can lead to higher reliability and act as a "regulatory device," and data seems to indicate the inclusion of a rubric has a positive effect on intra-rater reliability (Jonsson & Svingby, 2007, p. 136). Despite great care in rubric development, however, rubric rating can still prove difficult because raters, notwithstanding extensive training, will often internalize and apply the same rubric differently (McNamara, 1996; Eckes, 2011; Myford & Wolfe, 2003). Further, evidence suggests that raters don't always utilize the rubric as intended. For instance, Winke and Lim (2015) found that raters focused primarily on the left side column of the rubric and often completely avoided the "Mechanics" category to the far right.

To counter this problem, all raters must be trained and normed to have essentially the same mental picture of each of the rubric descriptors against which to compare examinee essays (Wolfe, 2005). Training and norming on rubrics has demonstrated a positive effect on inter-rater reliability (Eckes, 2011; McNamara, 1996). However, research repeatedly demonstrates that "raters typically remain far from functioning interchangeably even after extensive training sessions" (Eckes, 2011, p. 23). In a meta-analysis of 75 empirical studies, Jonsson and Svingby (2007) found that "agreement is improved by training, but training will probably never totally eliminate differences" of raters (p. 135). While rubrics offer an important mechanism by which rater reliability is bolstered, rater bias can prove to be disproportionately influential despite extensive training (Wilson & Case, 1997).

In response, sophisticated statistical measures, including the many-facet Rasch measurement model (MFRM), have been employed in the last three decades to compensate for rater effects (Engelhard, 1992; Eckes, 2011). Based on Georg Rasch's dichotomous measurement model (Eckes, 2011), MFRM was first proposed by Linacre in 1989 and is a linear model that essentially transforms observed ratings of performance tasks to a singular logit scale

based on successive log odds—probabilities a given examinee will receive a particular score by a specific rater (Eckes, 2011; McNamara, 1996). It is widely accepted as a fairly robust statistical mechanism that adjusts for rater effects and identifies outlying judges or examinees. The resulting fair average score is a modified score which corrects for rater bias and allows for more accurate assessment of ability.

MFRM has demonstrated increased inter-rater reliability through the mitigation of rater influence on examinee rating (Engelhard, 1992; McNamara, 1996). Proponents of MFRM point out its efficacy in accounting for rater effects—such as leniency or severity, central tendency, and halo (Myford & Wolfe, 2003)—and its versatility in addressing many potential sources of variability, or facets (Eckes, 2011). MFRM is one solution to the reliability issue in direct writing assessment, but it can be resource intensive and lead to reduced practicality.

2.2 Improving Practicality

Assessors responsible for placement or large-scale proficiency testing must operate within given financial and time constraints, so they often utilize practical indirect methods, such as multiple-choice tests, to assess a particular skill area. These methods, though reliable, are considered inadequate to the task of assessing writing (Greenberg, 1992). Therefore, the adoption of direct writing assessment represents a considerable improvement.

While rubrics enhance reliability, developing them can involve a lengthy process requiring collaboration among stakeholders and careful correlation with benchmarks for selected criteria (Brown, 2012). Further, the resources needed to train, calibrate, and ensure raters appropriately interpret a rubric can be both intensive and extensive, with trainings spanning anywhere from 3 or 4 hours to several days (Steedle & Ferrara, 2016). Training may require not

only instruction and practice in applying a rubric but also repeated norming, or calibration of raters, in order to maintain reliability (Office of Assessment of Teaching and Learning, 2016).

Although the addition of MFRM to traditional RR boosts rater reliability, some practicality may be sacrificed. MFRM calculations require additional time and resources not readily available to all, while the analyses necessitate specialized statistical software such as Facets and personnel with expertise to set up the data, program the execution files, and interpret the results (Linacre, 2017). Further, the cognitive loads associated with rubric use (Wolfe, 2005; Steedle & Ferrara, 2016) may have undesirable consequences: increased rating time per essay and a decreased number of essays that raters can reliably score before fatigue sets in (Ling, Mollaun, & Xi, 2014; Wilson & Case, 1997).

As practicality has become an increasingly important concern due to a culture of increased measurement in general and the growing need for standardized placement scoring, practitioners and program administrators continue to look for methods that effectively balance reliability and practicality without undermining validity.

2.3 Comparative Judgment

Rubrics and MFRM have reduced reliability issues while exacerbating practicality issues. An alternative to traditional rubrics and MFRM that purports to address both reliability and practicality is comparative judgment (CJ), which was first proposed by Thurstone (1927). It is based on the long-standing theory that humans are innately predisposed to successful comparisons but less apt when attempting absolute judgments in isolation (Laming, 2004; Fechner, 1980). Although the theory has been well researched in the field of psychology, CJ has only recently been used in education (Pollitt, 2004).

CJ provides an explicit frame of reference from which raters can judge the quality of the

work presented. When applied to direct writing assessment, raters compare two essays side by side and choose which essay is better (see Figure 1). Comparison data is aggregated and eventually rank-ordered on a scale representing the relative distance between each essay. The algorithms generating CJ data are based on well-established statistical models: the Rasch logistic model (Andrich, 1978) and the Bradley-Terry-Luce model (Turner & Firth, 2012).

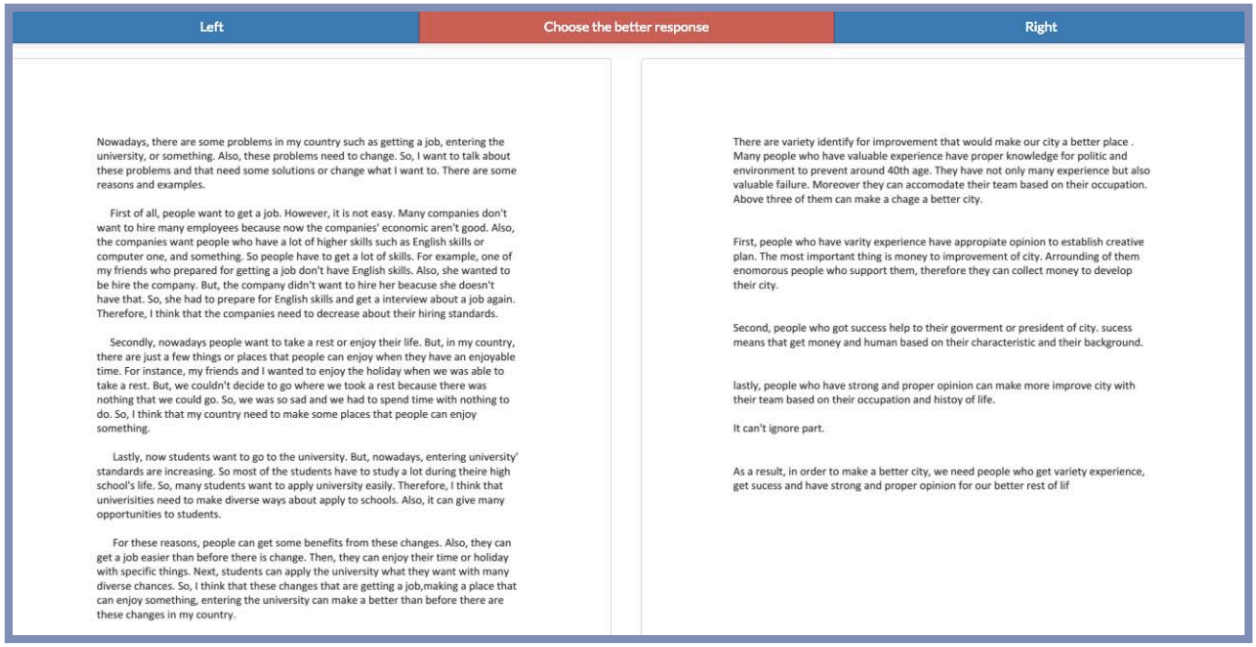

Figure 1. Example of two essays to compare in the CJ rating system used for this study.

Proponents argue the main advantage of CJ is its ability to estimate the subjective distance of objects that cannot otherwise be objectively arranged (Vasquez-Espinosa & Conners, 1982). This characteristic is particularly useful in the assessment of essays, which reflect real-world uses but can be difficult to assess reliably (O'Neill, 2011; Huot, 1990; Yancey, 1999). Further, the holistic nature of CJ assessment refocuses raters on the demonstration of skill without interference from rubrics, which can "create a barrier to the exercising of legitimate subjectivity by examiners" (Whitehouse & Pollitt, 2012, p. 3).

Jones and Inglis (2015) performed a multistage study in which they investigated the flexibility of CJ in addressing productive tasks by removing the "constraint for reliable marking"

(p. 352). Experienced test designers first developed a productive math assessment of higher order problem-solving skills for which experts then created what amounted to a cumbersome 16-page assessment rubric. After administering the exam to 750 students, it was marked traditionally by four experienced markers and judged using CJ by 20 math experts. Traditional marking resulted in high reliability, $r = 0.91$, and a strong correlation with predicted General Certificate of Secondary Education (GCSE) grades, $r = 0.73$; CJ reliability was $r = 0.86$ and correlated even more closely with GCSE predicted grades ($r = 0.76$). The same task resulted in comparable reliability and validity evidence using CJ without the need for such an unwieldy rubric, illustrating an advantage of CJ over RR in productive assessment design and rating.

Reliability, a possible advantage of CJ, may be bolstered by the method itself without compromising validity or practicality. "Probably the most immediate advantage of [CJ] is in making reliable the assessment of skills that are currently problematic" (Pollitt, 2012, p. 292). The forced comparisons inherent in CJ of themselves mitigate rater bias and problems typical in writing assessment, such as central and extreme tendencies or lenience and strictness; they are moderated by the very nature of the method (Pollitt, 2012; Bramley, 2007). Further, the algorithms underlying CJ can account for non-random missing data without affecting the results (Bramley, 2007) and eliminate the need for after-the-fact statistical modeling requiring trained personnel and expensive software. Steedle and Ferrara (2016) reported on multiple CJ studies in which all reliability indicators were > 0.73, and most were > 0.93. They further suggested that, due to the differences in RR and CJ, CJ may in fact be more valid and reliable than traditional rubric rating.

CJ shows promise as a more practical assessment method than MFRM as well. The amount of training required to perform CJ rating may be significantly less than what is necessary

for rubric rating (Jones & Wheadon, 2015). Pollitt (2012) reported on two studies, one involving children's writing and the other, e-portfolios, and indicated untrained raters were able to competently assess writing quality as both studies achieved reliability estimates of 0.96. Heldsinger and Humphry (2013) employed a form of CJ involving calibrated exemplars in which teacher ratings were compared with those of trained raters; they reported high inter-rater reliability for the teachers (.923) and a strong correlation between trained rater and teacher ratings (.895).

Comparative Judgment appears to excel in other measures of practicality as well. Raters report that CJ is less cognitively demanding and can be performed with greater ease (Christodolou, 2016) compared to traditional RR. In a CJ study involving 54 novice and experienced raters, 100% declared a preference for CJ over traditional marking, indicating it was less complex and refreshingly different (Pollitt, 2012).

Although promising, CJ is not without drawbacks. While uniquely suited to holistic assessment, it is cumbersome when applied to lists or other easily measured tasks or with longer tasks (McMahon & Jones, 2015). Further, CJ does not provide a feedback mechanism and is therefore unsuited for use in certain pedagogical contexts. The lack of rubric use, though positive in many ways, also removes the instrument by which a consensus of quality expectations is communicated (Brown, 2012). Finally, CJ relies on RR for the identification of benchmarks, which can be used as anchors in true score translations to the rubric scale, and therefore doesn't operate completely independently of RR.

Traditional CJ exhibits a further drawback in terms of practicality. Studies involving traditional CJ suggest anywhere from 25 to 50 judgments per essay are needed to assure quality results (Steedle & Ferrara, 2016; Whitehouse & Pollitt, 2012), as each essay needs to be

compared with almost every other essay in the set in order to accurately assign its place in the rank order (Vasquez-Espinosa & Conners, 1982). McMahon and Jones (2015) applied a more traditional CJ approach to teacher assessment: 20 judgments per examinee. Reliability and validity indicators of CJ were favorable: a reported 94.8% scoring consistency and 0.71 correlation with marking. Practicality evidence, however, was less favorable, as CJ was clearly more time-consuming than the single-marking method. The sheer number of comparisons can minimize CJ's usefulness as a rating method. Greater practicality, however, can be achieved with modernized forms of CJ made feasible through computer algorithms.

Adaptive comparative judgment (ACJ) can conceivably increase reliability and practicality simultaneously by leveraging initial Swiss rounds[1] to fuel an algorithm which generates the most informative pairings (Whitehouse & Pollitt, 2012). The predictive nature of ACJ minimizes the number of pairs required to complete the scale to as few as nine per examinee while maintaining reliability above .80 in most cases, an acceptable level for high-stakes testing (Wiliam, 2001; Whitehouse & Pollitt, 2012; Bramley, 2007; Pollitt, 2012; Heldsinger & Humphry, 2013; Steedle & Ferrara, 2016).

Critics question the efficacy of the adaptive nature of ACJ, however, arguing that preemptive pairings may lead to potential overinflation of ACJ reliability data (Bramley & Wheadon, 2015). Subsequently, a potentially more robust version with randomly distributed pairings has been developed (C. Wheadon, personal communication, August 5, 2017). This new version, which we will refer to as randomly distributed comparative judgment (RDCJ), also reduces the number of required judgments; however, unlike ACJ, which narrows the pairing

[1]Swiss rounds are typically used when the number of competitors makes the inclusion of all potential pairings infeasible. After an initial random round, subsequent rounds pair according to wins and losses and match pairs with similar scores.

possibilities as the algorithm advances, RDCJ maintains random pairings throughout the judging process, providing a more equitable judging framework for comparisons (Wheadon, 2015). These randomly distributed pairs, wherein each essay is judged an equal number of times and no two pairs are repeated, do not appear to overinflate reliability. Even with only nine judgments per examinee, RDCJ provides adequate data for the predictive algorithm to produce a reliable rank order congruous to those generated by the traditional CJ approach (Wheadon, 2015).

In a peer-assessment comparative study involving RDCJ and absolute measurement, the comparative condition outperformed the absolute condition in terms of reliability and validity evidence (Jones & Wheadon, 2015). Reliability measures for RDCJ were $r = 0.93$, $r = 0.82$, and $r = 0.85$ among the three schools involved with a combined correlation of 0.85 with expert scores. The absolute condition, in which raters directly assigned number scores without the added element of comparison, resulted in much lower reliability measures ($r = 0.28$, $r = 0.39$, and $r = 0.17$) and correlation with expert scores (0.07). Additionally, students were able to complete more RDCJ judgments than absolute judgments in the given time period, lending credence to its superior practicality, though exact rating times were not measured.

As a result, RDCJ shows potential as a more practical approach to direct writing assessment; yet to date there is little research comparing the practicality of RDCJ with traditional RR and MFRM. A comparison by Steedle and Ferrara (2016) of traditional RR with ACJ on direct writing assessment produced strong correlations between RR and ACJ scores on the two prompts used, 0.78 and 0.76. Reliability measures for ACJ were also high, exceeding 0.80 with a minimum of nine judgments. Recorded training times for both methods confirmed greater time requirements for RR over ACJ; however, reported rating mean times had to be estimated.

Another study by McMahon and Jones (2015) found a purely random form of CJ, without

the previously explained benefits inherent in the distributed approach, to be more time-consuming than single marking. However, again, results from this and other studies did not account for interruptions or time away from the computer during rating sessions (Whitehouse & Pollitt, 2012). Whitehouse and Pollitt (2012) call for a comparison of modernized CJ and traditional RR rating times, and Heldsinger and Humphry (2013) argue for further research "to understand the relative efficiency of the method in a range of educational contexts" (p. 233).

The broad applicability of CJ is still unexplored. There is not only a need for further investigation into the relative practicality of RDCJ compared to traditional RR and MFRM but also for a closer look into its applicability in relatively new contexts. "The question is no longer whether the method can work, but how widely could it, or should it, be used" (Pollitt, 2012, p. 281). Research so far has looked primarily at native English contexts, in areas ranging from peer assessment (Jones & Wheadon, 2015) to mathematical problem-solving (Jones, Swan, & Pollitt, 2014), but we can learn a lot from applying it to more complex scenarios such as L2 writing. Reports estimate the number of English-language learners will exceed two billion by the year 2020 (Beare, 2017), many of whom will be required to take direct writing assessments such as those in the Test of English as a Foreign Language (TOEFL) to demonstrate their proficiency.

Research directly comparing RDCJ, traditional RR, and MFRM, while strictly controlling for rating time, is needed to more fully investigate the relative practicality of RR, MFRM, and RDCJ in general, especially as they pertain to novel tasks and contexts. Given this need, we report on a study that assesses the validity, reliability, and practicality of RDCJ relative to the validity, reliability, and practicality of rubric rating with MFRM when applied to ESL writing.

The primary research questions were as follows:

- How do novice and experienced raters compare in terms of reliability when utilizing

traditional RR, MFRM, and RDCJ in an ESL setting?

- How do novice and experienced raters compare in terms of practicality when utilizing traditional RR, MFRM, and RDCJ in an ESL setting?
- How do novice and experienced rater scores compare with MFRM–validated scores in an ESL setting?

3. Method

This study used essays that had been double rated with a reliable rubric, analyzed with MFRM, and selected with fit statistics. Novice and experienced raters scored congruous sets of essays using RDCJ and RR with MFRM (see Figure 2).

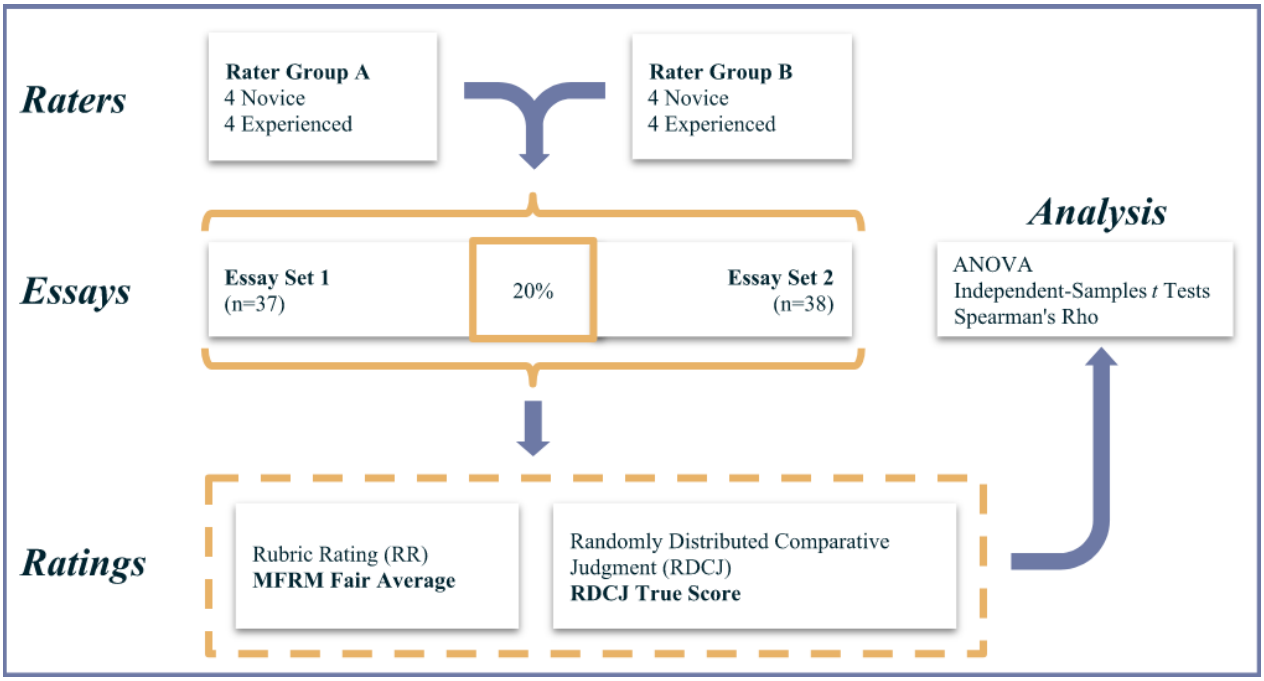

Figure 2. Study design to compare traditional rubric rating (RR) to many-facet Rasch modeling (MFRM) and randomly distributed comparative judgment (RDCJ). Analysis of variance (ANOVA) run to test for effects on rating time and Spearman's rho used to correlate between MFRM adjusted fair average, the study rubric rating fair averages, and RDCJ true scores to show evidence of validity.

3.1 Sample Essays

A stratified sample of 60 essays was selected from a pool of 30-minute ESL placement essays from an intensive English program (IEP). These MFRM rubric-rated essays were initially

rated by a group of experienced, rubric-trained raters who work as teachers at the IEP. They were further analyzed with MFRM and selected with fit statistics.[2] Efforts were made to not only select an even sampling from each of the rubric levels but also a representative sampling of the various language groups involved in the testing (see Table 1 in this section). To control for task, selected essays were collected from four prior rating sessions in which the same 30-minute essay prompt was given (see Appendix A). The strata are based on the rubric levels of 0 to 7, with 0 being little to no language or a reliance on simple, memorized words and phrases and 7 indicating university-level writing (see Appendix B), though no Level 7 essays were available to include in the study. The essays were further divided into two congruous sets of 37 and 38 essays, with 12 essays in common.

Table 1
Essay Levels and Language Background

	Essay Rating Levels						
Languages	0	1	2	3	4	5	6
Arabic	-	-	-	-	-	-	1
Chinese	-	1	2	-	-	-	-
French	-	-	-	-	1	1	-
Japanese	-	-	1	1	-	-	-
Korean	-	3	1	1	-	-	1
Mongolian	-	1	-	1	-	-	-
Portuguese	-	-	1	1	1	2	2
Russian	-	-	-	-	1	-	-
Spanish	2	3	5	5	5	6	6
Thai	-	-	-	-	2	1	-
Turkish	-	1	-	-	-	-	-
Totals	2	9	10	9	10	10	10

In order to have clear distinctions among the essay levels from which to compare the rating methods, care was taken to select essays that were typical models of each of the levels being tested. For rubric levels 1 to 5, selected essays represent a full-level difference between each rating level, based on the observed score, with 100% rater agreement of between 2 to 13

[2]Fit statistics are statistical measures which describe how closely assessment results align with expected outcomes based on existing patterns in the data.

raters per essay, and infit scores[3] between 0 and 1.36. Of the four essays rated "0," only two had a response to the 30-minute essay, so both of these were included in the study. Additionally, the small number of available Level 6 essays resulted in the application of slightly less rigorous selection standards and, as a consequence, a little less than a full-level difference in some instances. Fair average scores ranged from 5.62 to 6.67, and infit scores exceeded 1.5 on three of the essays (1.58, 2.35, and 3.22). Level 6 rater agreement, however, remained 100% with 2 raters per essay. Level 7 was not included because none of the available essays had been rated a 7.

3.2 Raters

There were two groups of evaluators (novice and experienced) who rated congruous sets of the essays using RDCJ and RR with MFRM. The novice and experienced raters were further divided into two groups. In order to measure method and order for novice and experienced raters, each group, A and B, was composed of four novice raters and four experienced raters.

The novice rater group consisted of eight raters selected from the Teaching English to Speakers of Other Languages (TESOL) undergraduate minor program at a large university in the western United States. This group was chosen in view of their preexisting interest in TESOL, as well as lack of previous knowledge and experience with RR. All novice raters were female undergraduate students in their early twenties.

The experienced rater group comprised eight individuals who self-selected from a pool of trained raters, either currently working or having previously worked as teachers and raters at the same IEP. These raters had from two to seven years of experience rating congruent placement essays with the study rubric in the same context. Of the eight raters, two were male and six were

[3]The ideal range for infit scores is .5 to 1.5. Anything over 1.5 indicates misfitting judges or examinees.

female; five were between the ages of 26 and 30, while the other three raters were between 36 and 45 years old.

3.3 Rating Methods

To attain acceptable levels of reliability, most rubric-rating systems employing MFRM use incomplete yet linked rating schedules (Eckes, 2011) in which at least two raters judge each essay. For the rubric-rating portion of this study, we employed a more conservative model in which each essay set was rated by all of the raters within the rating group to which it was assigned. We were also conservative when designing the CJ portion of the study. RDCJ has demonstrated acceptable levels of reliability,[4] with a minimum of nine judgments for each essay (Jones & Wheadon, 2015). Full essay sets were assigned to each of four distinct groups: Group A (novice), Group A (experienced), Group B (novice), Group B (experienced). For each set of 37 or 38 essays, raters were assigned an average of 100 judgments each (or approximately 11 judgments per essay).

3.4 Process

To control for order effects, each group began with a different rating mode. During the first rating session, Group A was assigned to RR Essay Set 1, while Group B used RDCJ to rate Essay Set 1. For the following rating session, they switched. Group A used RDCJ to rate Essay Set 2, and Group B was assigned to RR Essay Set 2.

Both groups received identical recorded instructions via video at the beginning of each rating session. Group members were instructed to complete the ratings independently in one session per rating method and avoid collaboration or discussion of essays with others in the study

[4]Reliability measures (Rasch person separation reliability) indicate the amount of error in the scores produced and are reported on a scale of 0 (a lot of error) to 1 (no error). Scores of .8 and above are considered highly reliable and acceptable for high-stakes testing.

until after the second rating session. Each rating session was completed electronically. For RR, an in-house rating system was used with some modifications to accommodate the more conservative design of this study. For RDCJ, a proprietary system called nomoremarking.com was used. Because the computer programs were designed to gather latency data, at the beginning of each rating session raters were instructed to log out of the system should they need to leave the computer for any length of time.

The RR session was preceded by a practice session, which included three practice ratings performed in the RR system. Each participant was instructed to use the paper copy of the rubric provided to complete the practice ratings (see Appendix B). Instructions for the RDCJ rating session included four practice essays (two comparisons) in the RDCJ rating system. Participants practiced choosing the "best" essay in each pair. Upon completion of the practice sessions, raters were instructed to begin rating.

Due to scheduling conflicts, the time and location for rating was adjusted for some raters. Novice raters in each of the groups attended rating sessions during on-site time slots with one exception: the final RDCJ session for Group A was performed remotely. Experienced raters completed their ratings off-site but were instructed to complete each distinct rating session in an uninterrupted time slot, with no outside collaboration until both sessions were completed. Researchers waited one week between rating sessions, thus allowing time to minimize recognition of the 20% of essays in common between both groups. This step was only cautionary, however, as the second session involved a fundamentally different rating method that would not likely be affected by the previous rating experience.

Consequent to the completion of the second rating session, all raters took a survey about their relevant background and experience data, as well as information relating to their rating

approach for each method, overall rating experience with both systems, and rating method preference (see Appendix C).

To examine traditional RR, the ratings of the 12 essays that were judged by all 16 raters were compared to the original MFRM ratings. The software program Facets was used to run an MFRM on the RR sessions from which measure scores, fair averages, infit statistics, and person separation reliability results were derived (Linacre, 2017). RDCJ data analysis was completed by a proprietary system operating from the nomoremarking.com website. Resulting downloads provided reliability scores, infit statistics, and true scores (or rankings) on a stochastic scale. To further assess the quality of the data, we used Spearman's rank-order correlation coefficient (Spearman's rho) to identify the level of linear association among the different variables. Additionally, we compared the initial raw scores with raw scores assigned by raters in this study in order to assess the impact of MFRM on study data. Finally, independent-samples *t* tests and an analysis of variance (ANOVA) were run to test for order, mode, and experience effects on rating time. Statistical analyses were calculated using SPSS® software, version 25.0 (SPSS Software).

4. Results

4.1 RR Raw Scores

Using the 12 essays that were rated by all 16 raters, we compared RR raw scores with the original RR raw scores according to rater background and essay level (see Table 2). While MFRM is robust with missing data and can be employed on incomplete rating schedules as long as there are some essays in common among all raters, this type of raw score comparison can only be conducted when the rating design is full-crossed and every rater rates every essay.

Table 2
A Comparison of Initial and Study RR Raw Scores

Original Rating Level	Essay	Experienced Raters											Novice Raters										
		0	1	2	3	4	5	6	7	Range	Exact Agreement (Original Rating)	Adjacent Agreement (Original Rating)	0	1	2	3	4	5	6	7	Range	Exact Agreement (Original Rating)	Adjacent Agreement (Original Rating)
0	35	2	6							2	25%	100%	6	2							2	75%	88%
1	28	2	5	1						3	63%	100%	7	1							2	13%	100%
	32		6	2						2	75%	100%	3	3	2						3	38%	100%
2	27			5	3					2	63%	100%		6	1	1					3	13%	100%
	31			5	3					2	63%	100%		5	3						2	38%	100%
3	36			6	2					2	25%	100%		3	4		1				3	0%	63%
4	34				1	6	1			3	75%	100%			3	5					2	0%	63%
	37					6	2			2	75%	100%			2	3	3				3	38%	75%
5	30					5	2	1		3	25%	100%			1	3	2	1	1		5	13%	50%
	33					3	3	2		3	38%	100%			3	1	1	3			4	38%	50%
6	26						1	7		2	88%	100%					1	2	3	2	4	38%	88%
	29						5	2	1	3	25%	100%			1			3	3	1	4	38%	88%
										2.42	53%	100%									3.08	28%	80%

= Original Rating

In looking at the relationship between novice and experienced raters, the disparity between novice rater agreement and experienced rater agreement is evident. Experienced raters exhibited 100% adjacent agreement on all of the essays in common, whereas 100% adjacent agreement was only present with four essays for novice raters. There were no instances of 100% exact agreement with either group. However, distinct differences lie between the groups in the exact agreement category as well. Experienced rater exact agreement exceeds 63% in seven instances and never falls below 25%; novice rater exact agreement, however, did not exceed 38%—with the exception of one instance of 75% exact agreement—and equaled 0% in two cases.

Patterns of severity and leniency are also evident. Weigle (1999), in a study investigating rater and prompt interactions using MFRM, found novice raters exhibited greater severity on graph essays than experienced raters, and that this difference disappeared after training. Data from this study supports this claim. Of the 69 novice ratings that were not exact, only 10 were higher, whereas 59 were lower. Experienced raters were more balanced: out of 45 ratings that were not exact, 22 were higher and 23 were lower.

These results suggest absolute rating without MFRM is less reliable. McNamara (1996) states "raw scores (the original ratings given by the judge) are no reliable guide to candidate ability" (p. 118). Despite clear superiority in rating ability exhibited by experienced raters, both groups, without MFRM, exhibit larger than acceptable variation in scoring. The low percentage of exact agreement of both experienced and novice rating groups, 53% and 28% respectively, means, at best, student placement based on this data would be accurate only half the time (with experienced raters) and, at worst, one quarter of the time (with novice raters). This level of variance does little to assure assessment quality and deliver fair results to examinees, especially when one examinee might happen to be rated by two severe judges while another would be rated by two lenient judges.

4.2 Reliability and Validity Estimates

Two separate measures of reliability were considered: reliability in terms of how reliably MFRM and CJ differentiate between the essays, and inter-rater reliability, or how reliably interchangeable the raters are. Validity evidence was primarily derived from a Spearman's rho correlation between (a) the initial MFRM-adjusted fair average, (b) the study rubric rating MFRM fair averages, and (c) the RDCJ true scores.

Reliability and validity estimates are presented in Table 3. Reliability reported is based on examinee separation: person separation reliability for RR and an analogous RDCJ reliability indicator (Wheadon, personal communication, August 5, 2017). Reliability measures ranged from 0.89 to 0.98 and tended to be slightly higher for rubric rating and experienced raters. Spearman's rank correlation coefficients involving the original MFRM fair average score were significant, ranging from 0.90 to 0.96, $p < .001$.

Table 3
Reliability and Validity Indicators

				Reliability	Validity
Group	Experience	Mode	*N*	Separation	rho
A	Novice	RR	36	0.96	0.94
		RDCJ	38	0.91	0.90
	Experienced	RR	36	0.98	0.95
		RDCJ	38	0.92	0.94
B	Novice	RR	37	0.96	0.96
		RDCJ	37	0.89	0.92
	Experienced	RR	37	0.96	0.94
		RDCJ	37	0.94	0.94

Note. RR=rubric rating; RDCJ=randomly distributed comparative judgment.

Although reliability data may not, of itself, speak to the quality of the data, it provides important information relative to the consistency of the data collected. Results indicate acceptable levels of reliability, above 0.80 (Nunnally & Bernstein, 1994), for both rating methods (MFRM and RDCJ) and rater backgrounds (novice and experienced). The reliability estimates were calculated four separate times: Group A (novice), Group A (experienced), Group B (novice), and Group B (experienced), revealing that within each of the aforementioned homogenous groups, raters of similar experience generated reproducible relative locations on a measure scale for the same essays. A further MFRM analysis was run comparing each of the mixed-rater groups, Group A and Group B. Data showed that rater reliability did not suffer as a result, as each reported a reliability estimate of 0.98.

Reliability indicators for both methods, though somewhat lower for RDCJ, were significant at .89 or above. Close reliability indicators within each method for Group A show no significant difference in rating background. Although experienced raters demonstrated slightly greater reliability than novice raters in Group B, the reliability of both groups exceeded industry standards for high-stakes testing. Therefore, novice and experienced raters were essentially interchangeable with regards to reliability when employing either the fully crossed RR with MFRM or RDCJ.

A Spearman's rank-order correlation coefficient involving the original MFRM fair average score provides correlational indicators of data quality. Results echo that of reliability as again no significant difference is evident between novice and experienced raters and indicate that from 92% to 96% of scores generated by participating raters, novice and experienced, can be explained by the original writing placement score. Peripheral validity evidence that correlates rating scores of the study with level placement is also significant, as anywhere from 86% to 94% of the novice and experienced rater scores can be explained by final-level placement in the IEP.

Spearman's rank-order correlation coefficient provided evidence of the concurrent validity, or quality, of the scores resulting from this study. Novice RR demonstrated slightly higher correlations with the original scores than novice RDCJ, whereas experienced RR and RDCJ correlations with the original scores were essentially the same. However, in every case the data was highly correlated, $r_s > .90$, $p < .01$, suggesting that both novice and experienced raters, utilizing either method, were essentially synonymous.

Raw score rating disparity between experienced and novice raters supports what others have found, that training minimizes rater effects. However, the remaining gap between experienced rater raw scores and initial ratings points to the need for statistical procedures that

account for rater bias. MFRM, under heightened coverage conditions of eight raters per examinee, essentially eliminated the distinction between the two groups, demonstrating the strength of the model in accounting for rater effect. Under typical double-rating conditions, however, it is unlikely that MFRM would be able to model out all of the rater variance in novice raters. RDCJ is not subject to the same limitations; it achieved comparable validity and reliability for both novice and experienced raters while operating within more typical rating parameters.

4.3 Practicality

Practicality was measured in terms of mean and median rating times per essay (see Table 4) according to rating method, rater experience, and rating session order (first session, second session). Mean time comparisons according to experience, method, and order appear in Table 4. The table also presents Cohen's *d*, which is an effect size that illustrates the standardized differences between the pair-wise means.

Table 4
Comparisons of Mean Time According to Experience, Method, and Order

					Cohen's *d*							
Group	Experience	*N*	*M*	*SD*	1	2	3	4	5	6	7	8
A-RR	1. Novice	36	92.9	52.7		0.32	0.95	0.82	2.32	2.05	1.95	2.05
	2. Experienced	36	79.2	31.2	-0.32		0.87	0.78	3.23	2.77	2.59	2.78
B-RR	3. Novice	37	52.8	28.7	-0.95	-0.87		-0.15	2.26	1.78	1.59	1.79
	4. Experienced	37	56.9	25.3	-0.82	-0.78	0.15		2.80	2.22	2.01	2.23
A-RDCJ	5. Novice	38	6.4	3.7	-2.32	-3.23	-2.26	-2.80		-1.72	-2.15	-1.46
	6. Experienced	38	15.7	6.8	-2.05	-2.77	-1.78	-2.22	1.72		-0.50	0.08
B-RDCJ	7. Novice	37	19.4	7.7	-1.95	-2.59	-1.59	-2.01	2.15	0.50		0.55
	8. Experienced	37	15.2	7.7	-2.05	-2.78	-1.79	-2.23	1.46	-0.08	-0.55	

Note. N=number of essays. M=mean time in seconds. SD=standard deviation.

These results may suggest an order effect. Group A raters performed RR during their first rating session, whereas Group B raters completed RR during their second session. An effect size of 0.95 between both novice RR groups is very strong. A pairwise comparison of both Group A and Group B's experienced RR times produced a similarly strong effect size of 0.78. The strongest effect size in this category, 2.15, was between Group B and Group A novice RDCJ

ratings, and the weakest effect size overall was between both experienced RDCJ groups. This effect size, 0.08, demonstrates the RDCJ rating time difference between the two groups was trivial.

Background, with one exception, also appeared to affect rating times. Within Group A, the novice to experienced RR effect size of .32 is small yet nontrivial. The RDCJ rating time effect size comparing experienced to novice raters for the same group demonstrated an effect size of 1.72. Within Group B, the experienced to novice RR effect size of 0.15 was trivial; Group B novice to experienced RDCJ, however, had a medium effect size of 0.55.

Method had the greatest impact on rating time. Within Group A, the novice RR to novice RDCJ rating time effect size was strong at 2.32. In the same group, the experienced RR to experienced RDCJ rating time effect size was even larger at 2.77. Within Group B, the novice RR to novice RDCJ was also large, though less marked, at 1.59. The Group B experienced RR to experienced RDCJ large effect size of 2.23 add further support for the strength of the effect of method on rating time.

Table 5 presents results of three independent-samples *t* tests[5] comparing the effects of order, background, and method on rating times per rating decision. In terms of an order effect on rating times, an independent-samples *t* test comparing the mean times of the groups who RR first with the groups who RR second found a significant difference between the means of the two groups, t(3764) = -4.78, p = .000. The overall mean of the groups who performed RR first (M = 28.8, *sd* = 42.3) was significantly lower than the mean of the groups who performed RR second. An independent-samples *t* test comparing the effect of background on rating time was also significant. The mean rating time comparison of both novice and experienced raters revealed a

[5]For all three T-tests the Levene's test was significant, so equal variances are not assumed.

significant difference between the means of these two groups, t(3749.6) = -2.58, p = .000. The mean time of the novice group was significantly lower than the mean time of the experienced group (MD = 3.6). Method was also significant and revealed a greater mean difference in rating times between RR and RDCJ (MD = 53.9). This final independent-samples *t* test, comparing the mean times of RR and RDCJ, found a significant difference between the two group means, t(641.7) = 20.6, p = .000.

Table 5
Mean Essay Evaluation in Seconds According to Method, Background, and Order

		Mean Time (in seconds)	*t* value	df	SD	*p* value (2-tailed)
Method	*Rubric Rating*	77.7	20.6	641.7	61.6	.000
	RDCJ	23.8			32.0	
Background	*Novice*	30.3	-2.58	3749.6	41.2	.000
	Experienced	33.9			44.3	
Order	*First Session*	28.8	-4.78	3764.0	42.3	.000
	Second Session	35.4			43.0	

*Note.*SD=standard deviation. RDCJ=randomly distributed comparative judgment. df=degrees of freedom.

A between-subjects ANOVA was calculated to examine the effect of method and background on rating time while controlling for order. Order was significantly related to rating time, $F(1, 3761) = 28.11$, p = .000. The main effect for background was significant, $F(1, 3761) = 3.92$, p = .048, with novice ratings that were significantly faster (M = 30.31, *sd* = 41.19) than experienced ratings (M = 33.91, s*d* = 44.29). The main effect for method was also significant, F(1, 3761) = 995.06, p = .000, with RDCJ rating times that were significantly faster (M = 23.77, *sd* = 31.98) than RR times (M = 77.67, *sd* = 61.65).

After accounting for the effect of order on rating mean times, ANOVA results provide further evidence of a method effect on rating time. With an overall mean rating time difference of 53.9 seconds between RR and RDCJ, the data represents a 226% increase in the time it takes for RR over RDCJ. Background and order, though significant, did not demonstrate the same level of influence on rating time with mean differences of 3.6 and 6.6 seconds, respectively.

Median rating times according to experience and method are shown in Figure 3. The median time to complete each rating decision was significantly greater for RR than RDCJ. Experience was not a significant factor. For Group A, novice raters took less time for RDCJ but more time for RR. For Group B, experience had little effect on the median rating time. Additionally, a potential learning effect resulted in lower times for the second RR session but not for RDCJ.

Method demonstrated the greatest effect on rating time. All RDCJ rating session median times were significantly different from the RR session median times, though, with the exception of the Novice A RDCJ rating session, not significantly different from each other. These results are indicative of a method effect on rating time. Median times for RDCJ ratings per essay ranged from 5.6 to 17.6 seconds, a difference of 12 seconds. The difference between the lowest RR median time (52.5) and the highest RDCJ median time (17.6), however, is substantially higher at 34.9 seconds. According to these numbers, it would take between 9.33 and 29.33 minutes per rating decision for 100 ESL essays using RDCJ, whereas it would take substantially longer, between 1.46 and 2.54 hours, per rating decision using RR with the same number of essays.

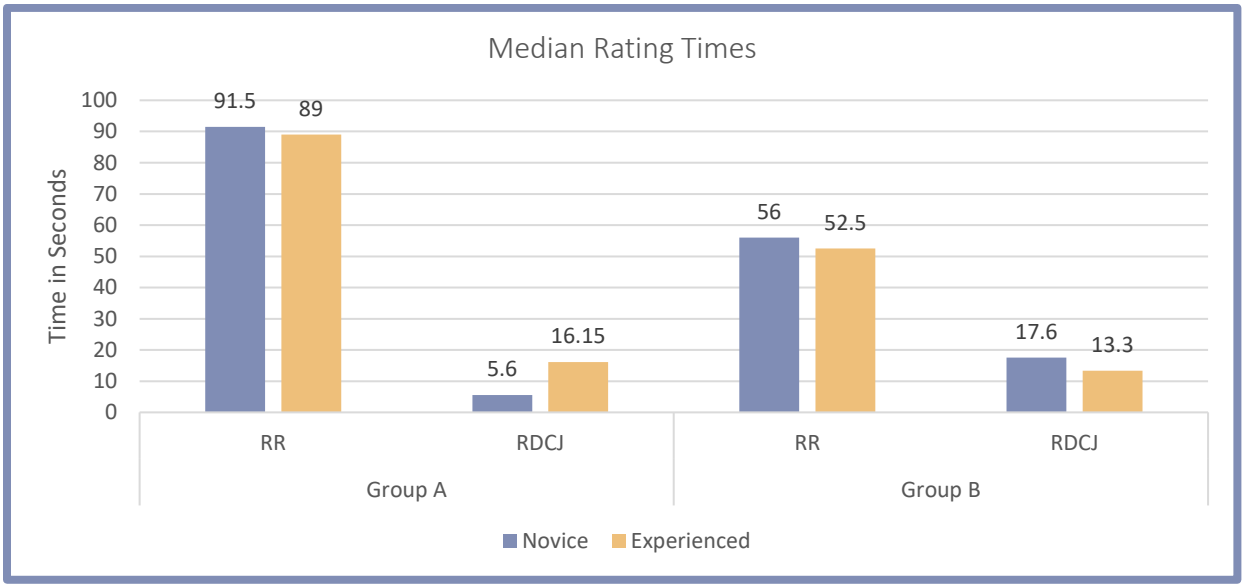

Figure 3. Median time according to method and experience to complete the rating assessment for RR and for RDCJ.

Median time improvement according to background and method is shown in Figure 4. Improvement in time per rating decision was greatest for RR, especially for novice raters, who lowered the median time per essay by 35.5 seconds between the first and second RR sessions. Experienced rubric raters lowered their time by 36.5 seconds. Novice raters for RDCJ demonstrated a 12-second improvement, whereas experienced RDCJ raters showed no significant improvement. Novice raters, overall, showed the greatest improvement in rating time between the first and second sessions, regardless of method. Experienced rater time improvement was evident for RR but essentially nonexistent for RDCJ.

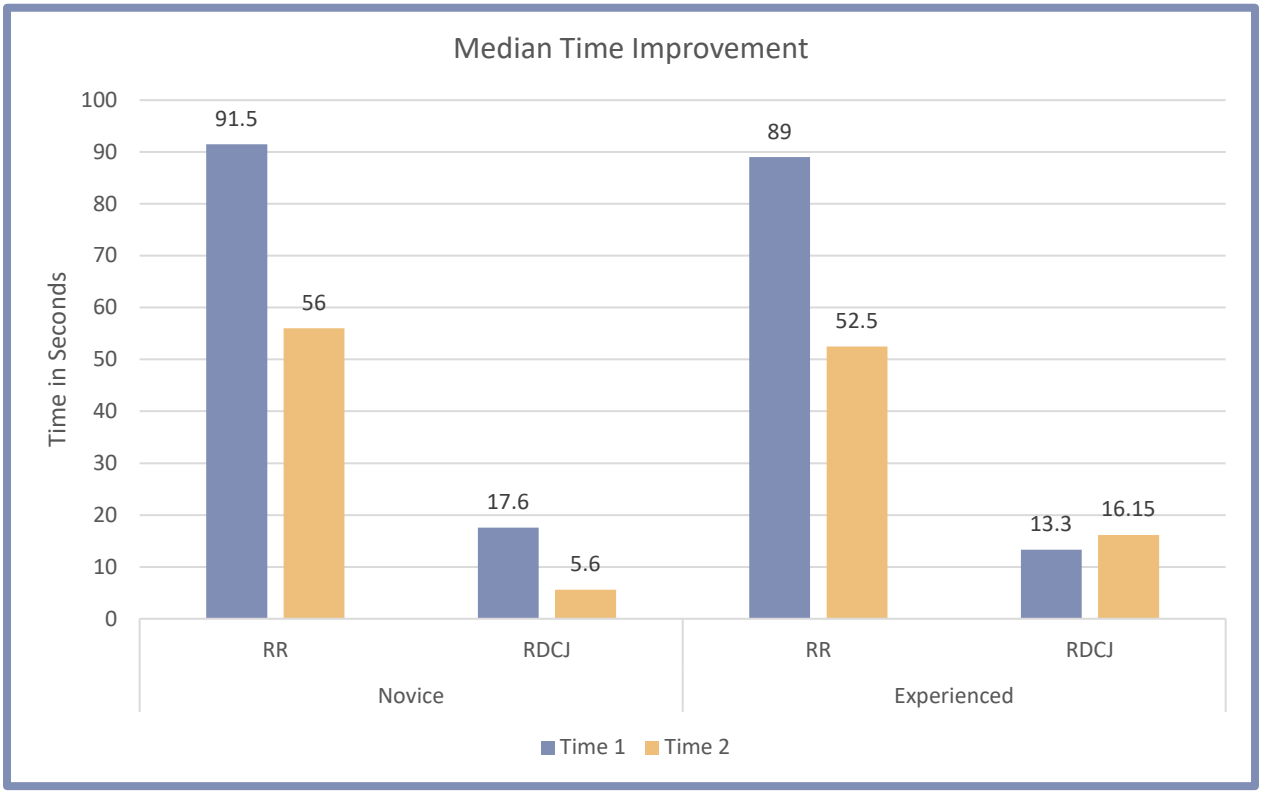

Figure 4. Mean improvement in time according to experience and method for RR and RDCJ assessments.

Study data indicates significant variance in rating time according to method. All RDCJ rating sessions presented significantly faster rating times than the rubric rating sessions, regardless of rater background. The fully crossed RR model used in this study potentially

overinflated RR mean times. However, even when halving the RR average median times, differences indicate RR takes 23 seconds longer per judgment than RDCJ.

Order effect provided evidence of increased rating speed for subsequent rating sessions. It is unclear what caused this effect, but research suggests that rater drift, a phenomenon in which rater efficacy diminishes over time, may be a contributing factor. In a study on rater performance over time, Myford and Wolfe (2009) discovered that rater performance is not devoid of error and bias and may change over time.

It is important to note, however, that mean and median rating times, though helpful indicators, are not representative of final rating times. Varying rating schedules in RR, or the assignment of number of ratings per examinee in RDCJ, will affect timing results. The information provided here can act as a guideline as assessors seek to balance reliability, which increases as the number of judgments per examinee increases, and practicality, which decreases as the number of judgments per examinee increases. Final rating times might vary widely depending on the chosen structure of the rating schedules.

In addition to rating time, we collected data relative to rater experience and training. Training and norming of raters can require a substantial time investment (Tarricone & Newhouse, 2016). Trained raters reported an average of 3 years of rating experience at the IEP, totaling about 144 hours of training and norming sessions. In addition to the time investment, the approximate cost to pay for training and norming sessions of the experienced raters in this study is $3,200, which doesn't include the cost or time investment on the part of rater trainers or other pertinent personnel and resources.

A final area of consideration was rater preference. The rubric model requires raters to maintain a mental image of each point along the rating scale for comparison with the examinee,

whereas CJ makes the necessary comparisons explicit by presenting both items side by side. Rubric rating can be highly cognitively demanding (Wolfe, 2005), and perhaps unnecessarily so. However, the relative tedium with which raters may approach absolute rating when compared to comparative rating may not only impair their enjoyment of the task but perhaps narrow the length of time they can viably perform it (Ling, Mollaun, & Xi, 2014). Christodoulou (2016) pointed out that rater response to this new method was strongly positive and reported the following rater comments regarding the use of CJ: "quicker and speedier than traditional moderation," "easier and less taxing to make judgments," and "results did feel intuitively right." Other studies have found that CJ judgments were faster and easier for judges and less cognitively demanding (Steedle & Ferrara, 2016).

In a post-study survey (see Appendix C), we collected data related to rater preference. Out of 16 raters, 12 indicated a preference for CJ over RR. Three main categories emerged in their comments: (1) CJ was faster, (2) CJ was easier, and (3) they were more confident in their decisions. They made comments like, "I think my answers were closer to being accurate with comparative. I also enjoyed it a lot more." In reference to CJ, others provided comments such as: "So much faster. So much less to think about. The rubric is often a little intimidating. I generally felt more confident," and "Lower learning curve, easier to simply compare two essays and choose which is better. Also much faster."

Interestingly, of the four raters who preferred RR over RDCJ, three were novice raters. Those who preferred RR made comments like, "I am more familiar with it. … That's why it's easier for me." "Even though it's harder to do this system and I want more training on it before doing it again, I think Rubric Rating is more accurate." "I'm a bit concrete sequential. I like order. I like steps. … That is something I like about rubric rating. There is a scale." It is possible

the preponderance of experienced raters who are pro-RDCJ is in part due to the novelty of the new method.

5. Limitations

Results of the MFRM analysis involving novice raters were somewhat surprising considering the given irregularities in their ratings. It is likely the atypical fully crossed rubric rating model, as well as the carefully selected stratified essays utilized in the study, overinflated the RR reliability indicators to some degree. MFRM is not the panacea it may appear to be at first glance. In the real world, such a scenario would likely never happen. A modified rating schedule with more typical rating samples would reasonably achieve different results.

Survey comments indicated rater concern over the possibility that essay length acted as a proxy indicator of quality. Comments included the following: “I had to try hard not to immediately mark the longer or better-formatted essay as the winner.” “I tried not to let the length make me biased but it was hard.” “Most of the time, longer ones with distinct paragraphs were rated higher. If they were even in those regards, I then skimmed through the text for language.” We suggest this as an area for further research.

There was a notable anomaly in the RDCJ rating time data. The mean rating times for three of the four RDCJ groups (Group A experienced, Group B novice, and Group B experienced) were all within a similar range: 19.4 seconds, 15.7 seconds, and 15.2 seconds, respectively. The Group A novice RDCJ mean rating time of 6.4 seconds may be the result of variables not accounted for in this study. As such, its inclusion in the study may have affected the accuracy of the results.

6. Conclusions

Several important conclusions can be drawn from this study. While training was shown to increase reliability in RR, it was still inadequate. However, MFRM under typical double-rating conditions would also likely be unequal to the task of modeling out all novice rater variations. Under typical conditions, RDCJ, however, produced analogous results regardless of background and proved appropriate for use in an ESL direct writing assessment context.

Similar reliability and validity evidence between experienced raters employing MFRM and experienced or novice raters using RDCJ spotlights practicality as a point of discrimination. The mean difference in rating time, the additional time and resources required for rater training and norming, and rater preference clearly point to RDCJ as a more practical rating method. Programs operating within budgetary and time constraints, as well as large-scale testing organizations, may find RDCJ a viable alternative for placement and proficiency testing.

However, whereas RR with MFRM automatically incorporates the rating scale in data reports, RDCJ true score conversion requires the inclusion of examinee essays already rated using RR to act as anchors with which the relationship of new examinees can be measured. So, in essence, RDCJ is partially dependent on RR to translate scores into real-world contexts. Perhaps a marriage of the two methods will capitalize on the strengths of each: smaller numbers of highly trained raters can identify benchmarks to anchor RDCJ while larger numbers of less experienced raters can perform RDCJ on the bulk of the examinees, delivering quality, reliable, and practical results that are meaningful in the given context.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2*(3), 449–460.

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.

Beare, K. (2017, May). Retrieved August 4, 2017. How many people learn English? *ThoughtCo.* Retrieved from https://www.thoughtco.com/how-many-people-learn-english-globally-1210367

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

Bramley, T., & Wheadon, C. (2015, November). The reliability of Adaptive Comparative Judgment. *AEA-Europe Annual Conference.* Cambridge, UK: Cambridge Assessment. Retrieved from http://www.cambridgeassessment.org.uk/Images/296241-the-reliability-of-adaptive-comparative-judgment.pdf

Breland, H. M. (1983). *The direct assessment of writing skill: A measurement review* (College Board Report No. 83.6; ETS RR No. 83.82).

Brown, J. (2012). *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages.* Honolulu, HI: National Foreign Language Resource Center.

Christodoulou, D. (2016, September 15). Research ED [Video file]. Retrieved from https://www.youtube.com/watch?v=3tpUBbLK_g8

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

Effect Size Calculator for T-Test. (n.d.). Retrieved February 10. 2018, from http://www.socscistatistics.com/effectsize/Default3.aspx

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*(3), 171–191.

Fechner, G. T. (1980). *Elements of psychophysics* (H. E. Adler, Trans. Howes & Boring, Eds.). New York: Holt, Rinehart, & Winston.

Greenberg, K. L. (1992). Validity and reliability issues in the direct assessment of writing. *WPA: Writing Program Administration, 16*(1–2), 7–22.

Heldsinger, S. A., & Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educational Research, 55*(3), 219–235.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201–213.

Huot, B., O'Neill, P., & Moore, C. (2010). A useable past for writing assessment. *College English, 72*(5), 495–517.

IBM SPSS Statistics Software (Version 25) [Computer software]. Armonk, NY: IBM Corp.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774–87.

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics, 89*(3), 337–55.

Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*(1), 151–77.

Jones, I., & Wheadon, C. (2015) Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47,* 93–101.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2,* 130–144.

Laming, D. (2004). *Human judgment: The eye of the beholder.* London, UK: Thomson.

Linacre, J. M. (2017). Facets computer program for many-facet Rasch measurement (Version 3.80.0) [Software]. Beaverton, Oregon: Winsteps.com

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Education Testing Service, 31*(4), 479–499.

McMahon, S., & Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice, 3,* 368–389.

McNamara, T. F. (1996). *Measuring second language performance.* New York, United States: Longman.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–49.

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

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371–89.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Office of Assessment of Teaching and Learning. (2016). Retrieved Aug. 7, 2017. *Quick guide to norming on student work for program-level assessment.* Retrieved from Washington State University website: https://atl.wsu.edu/documents/2015/03/rubrics-norming.pdf

O'Neill, P. (2011). Reframing reliability for writing assessment*. The Journal of Writing Assessment, 4*(1), n.p.

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

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300.

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

Tarricone, P., & Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Education, 13*(16), 1–11.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–86.

Turner, H., & Firth, D. (2012). Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software, 48(9),* 1–21.

Vasquez-Espinosa, R. E., & Conners, R. W. (1982). *The law of comparative judgment: Theory and implementation* (RSIP TR 403.82). Baton Rouge, LA: Louisiana State University. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a136169.pdf

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145–78.

Wheadon, C. (2015, February 10). Retrieved Aug. 7, 2017. *The opposite of adaptivity?* [Blog]. Retrieved from https://blog.nomoremarking.com/the-opposite-of-adaptivity-c26771d21d50

Whitehouse, C., & Pollitt, A. (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment.* Centre for Education Research and Policy. Manchester, UK.

Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3–13, 29(3),* 17–21.

Wilson, M., & Case, H. (1997). *An examination of variation in rater severity over time: A study in rater drift* (Berkeley Evaluation and Assessment Research Center report). UC Berkley.

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing, 25,* 37–53.

Wolfe, E. W. (2005). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment, 2*(1), 37–56.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50*(3), 483–503.

Appendix A. 30-Minute Essay Prompt

Identify one improvement that would make your city a better place to live for people your age and explain why people your age would benefit from this change. Use specific reasons and examples to support your opinion and describe the potential immediate and long-term consequences of this improvement. You have 30 minutes to write your response.

Appendix B. ELC Writing Rubric

Level	Text Type	Content	Accuracy
	- Length - Organization	- Functional Ability with the Language - Vocabulary	- Grammatical Complexity - Meaning
7	Essays may be a full page or more. Organization and transitions make writing very easy to read and understand.	Able to write more complex elaborations (i.e., summaries and paraphrases dependent on task). Uses a range of general and academic vocabulary. Writing uses a variety of cohesive devices. Provides sufficient background information as evidence that the writer is generally aware of the readers' needs. Readily understood by native readers.	Excellent control of a full range of grammatical structures. Small errors in grammar, syntax, spelling, or punctuation may occasionally distract a native reader, but there is no evidence of a pattern of errors. Writing is easy to read, but the writer may fail to convey the subtlety and nuances of the language.
6	Multiple paragraph essays with clear organization.	Appropriately uses abstract and concrete language to convey meaning. Message is pragmatically accurate for easy reading. Attempts to use cohesive devices, but they may be redundant. Wide and varied general and academic vocabulary and topics.	Able to use language in detail in all time frames. Control of syntax in word order, coordination, and subordination, while not perfect, does not distract greatly from meaning. No or very few spelling problems. Evident use of a wide range of structures. May be a few errors with complex and infrequent grammatical structures.
5	Multiple paragraphs with evidence of organizational markers on the essay level.	Able to meet all practical writing needs. Favors concrete ideas, and some more abstract topics may be discussed, but meaning is perhaps unclear. Vocabulary is quite varied, but not to the extent of level 6.	Able to use language in major time frames. There is apparent subordination, but it is more like oral discourse. Mastery of grammar with simple sentences. More complex sentences are attempted but contain errors and may not be clear.
4	Multiple paragraphs are present with organization on the paragraph level (topic sentence, supporting detail, etc.) but perhaps not on the essay level.	Writing is usually in the context of personal interests and experiences, daily routines, common events, and immediate surroundings. Concrete topics are discussed. Some examples and explanations may not be clear. Some points may not be well supported or explained.	Some mastery of past narration (past progressive, simple past, etc.) with both regular and irregular verbs. Inconsistencies occur in other time frames. The majority of sentences will be shorter. Complex sentences are common and generally accurate. Problems in accuracy may occur, and the overall meaning may occasionally be obscured.
3	At least one paragraph (for 30-minute writing portion). Organization is weak with multiple paragraphs.	Able to meet some limited practical writing needs—writing about personal interests and experiences, daily routines, common events and immediate surroundings. Structure and meaning are highly predictable. Usually relating to personal information or immediate surroundings. Writing exhibits a small range of vocabulary.	Solid writing of short and simple conversational-style sentences with basic subject-verb-object word order. Exhibits some consistent success with compound and complex sentences. Basic errors in grammar, word choice, punctuation, and spelling. Most writing framed in the present. Some mastery of past narration in the simple past with regular verbs. Other time frames may be attempted with some success. However, natives used to the writing of nonnatives can usually understand the meaning.
2	Simple sentences: some compound and complex sentences with repetitive structure. Lacks clear paragraph organization.	Close, personal explanations with very limited vocabulary. Writers can express themselves within a very limited context.	Able to write clear simple and compound sentences with limited vocabulary and conjunctions. Attempt to create some compound sentences using connectors like "because." Writing is successful in present tense, occasional, and often incorrect use of past or future tenses. Text is writer-centered.
1	Some simple sentences.	Reliance on formulaic or memorized language.	Exhibit accuracy when writing on well-practiced familiar topics using limited formulaic language. Sentence-level constructions. The volume of writing may be so small that it undermines the reader's ability to evaluate accuracy, or errors occur so frequently that the purpose of the writing task may not be completely clear.
0	Able to supply limited information on forms and documents (i.e., names, numbers, nationality, etc.).	With adequate time and cues, may be able to produce a limited number of isolated words.	Inability to use sentence forms. Volume of writing is insufficient to assess accuracy.

Appendix C. Pertinent Survey Questions*

What is your name?
What is your gender?
How old are you?
Are you a native English speaker?
What is your current education level?
Have you received rater training?
When was your most recent rater training?
Have you received writing rater training at the IEP?
When did you receive your *first* writing rater training?
When did you receive your *most recent* writing rater training?
In total, how long did you or have you worked as a writing rater at the IEP?
Have you rubric-rated writing in another context other than the IEP placement exams?
Describe your rubric rating training and rating experience in the context other than the IEP placement exams.
Prior to this study, have you rated using comparative judgment?
Describe your prior experience using comparative judgment rating.
How difficult were each of the rating modes? [likert scale]
How accurate did you feel your ratings were with each rating mode? [likert scale]
How fast were you able to make rating decisions with each rating mode? [likert scale]
Describe the process you used for making rating decisions using the rubric rating method.
Describe the process you used for making rating decisions using the comparative judgment method.
In your opinion, what are some of the pros of rubric rating?
In your opinion, what are some of the cons of rubric rating?
In your opinion, what are some of the pros of comparative judgment?
In your opinion, what are some of the cons of comparative judgment?
As a rater, which method do you prefer?
Why do you prefer [method selected]?

*Did not include all questions, as not all questions were relevant.