Fall 2018

# Evaluating rater effects in the context of ethical reasoning essay assessment: An application of the many-facets rasch measurement model

Madison A. Holzman
*James Madison University*

# Evaluating rater effects in the context of ethical reasoning essay assessment: An application of the Many-Facets Rasch Measurement Model

Madison A. Holzman

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Department of Graduate Psychology

December 2018

FACULTY COMMITTEE:

Committee Chair: S. Jeanne Horst

Committee Members:

Allison Ames

Christine DeMars

John Hathcoat

Bill Hawk

Lori Pyle

**Acknowledgements**

First, I would like to thank Dr. Jeanne Horst, my advisor through this process. You provided incredible support my entire time in the program. You're a wonderful mentor and I admire your passion and quiet leadership. I have learned so much from you these past few years, and know I will continue to learn from you in the years to come. Thank you for everything.

Second, I would like to thank my committee: Dr. Allison Ames, Dr. Christine DeMars, Dr. John Hathcoat, Dr. Bill Hawk, and Dr. Lori Pyle. Allison, I feel so glad to have had the opportunity to work with you. You've been a role model for me since starting the program, and I value the time we had to work together. Christine, you are one of the sweetest people I know. You're also one of the smartest people I know – I hope one day I can have thoughts half as insightful as you've provided me. John, I see the field differently from having worked with you. You've taught me to question assumptions and to always be curious. Bill, I've learned so much from working with you. I've also loved revisiting my philosophy days during our meetings –you've helped me fill a void that I've missed since undergrad. Lori, I can't overstate how much I look up to you. I learn from you in every meeting, and I am so glad I've had the opportunity to work with you. I'm also glad you trust me with Karl Kitten and let me have some pet time.

Next, I would like to thank Ethical Reasoning in Action: The Madison Collaborative. You not only financially supported this project, but were excellent partners along the way. Your thoughts and feedback were instrumental in the success of this project, and I hope you gained as much from this work as I've gained from working with you.

Of course, I have to thank my CARS and A&M family. To the students past and current, I wouldn't have made it through without you. Though late nights and early mornings aren't the best, they were better with all of you. I've learned so much about our field and about myself from having worked with you all, and I am honored to call you my colleagues. To the faculty, thank you for your support these past few years. When I was meeting with you to see if this program was something I wanted to pursue, I knew there was something special here, and I'm fortunate to have had to opportunity to work with each of you and call you my colleagues.

Finally, I would like to thank my friends, family, and friends who are family. Aaron, Carson, Danae, Eileen, Jamie, Sarah, and Thai, you've been the best support system I could ask for. Thank you for the late night food runs, sitting with me even when I was too tired to talk, the surprise snacks, and making me take breaks to go on adventures. I can never repay you for all you've done for me. I don't know what I did to make friends like you along the way, but I'm lucky to call you my people. To my friends at Shenandoah Valley Performance Clinic, thank you for giving me a space where I could leave school and work at the door. Even on the toughest days, I found peace and confidence on the gym floor. Thank you for creating that space. To my family, thank you for the support. It's been years of phone tag and barely seeing each other, but I promise to visit more often.

## Table of Contents

## List of Tables

## List of Figures

## Abstract

Performance assessments are an often desired type of assessment due to their potential for alignment between the assessment and reality. However, due to the rater-mediated nature of scoring (Eckes, 2015), performance assessments have psychometric challenges that cannot be ignored in testing and assessment work. Specifically, performance assessment scores are prone to rater effects, or systematic differences in how raters evaluate performance assessment products (Myford & Wolfe, 2003). The purpose of this project was to evaluate ethical reasoning essay scores for rater effects. The Many-Facets Rasch Measurement (MFRM) model was used to evaluate ethical reasoning essay scores for rater leniency/severity effects, restriction of range, and rater leniency/severity by rubric element interaction effects. Individual rater leniency/severity effects were observed in this sample of raters, as was an interaction effect between rater leniency/severity and rubric element. Moreover, a restriction of range effect was observed, with scores restricted primarily to the lower end of the rubric score categories. To provide a preliminary explanation for differences in rater leniency/severity, the relationship between raters' knowledge of ethical reasoning and their leniency/severity was evaluated. No relationship between raters' knowledge of ethical reasoning and their leniency/severity was observed in this study. Based on findings, recommendations are made for rater training. Specifically, ethical reasoning program coordinators may consider using the MFRM analysis during rating to identify individual raters who are exhibiting rater effects. Program coordinators may then work with individual raters on additional training and rubric calibration to mitigate individual rater effects. Additionally, recommendations are made regarding the statistical adjustment of student scores to

mitigate rater leniency/severity effects in the ethical reasoning scores. Though score adjustment is attractive if the goal is to mitigate rater leniency/severity effects, it has implications for inferences made from scores. Future research may focus on further identifying causes of rater effects, as well as methods for mitigating rater effects.

# Chapter 1: Introduction

Internal and external accountability calls require educators to demonstrate that students meet academic degree program and institutional learning outcomes. That is, educators must assess whether students are learning the knowledge and skills deemed critical by educators and key stakeholders. The manner in which students' knowledge and skills are assessed may vary based on student learning objectives. Content-based outcomes may be assessed through selected-response (i.e. multiple-choice, matching, true-false) assessments, on which students select the appropriate response from a multiple-choice list or match responses to draw connections between ideas. Other objectives may be better assessed by asking students to produce a product or engage in a process, also known as performance assessment (Johnson, Penny, & Gordon, 2009).

Higher education is currently in the midst of a push for the use of performance assessments to evaluate collegiate student learning objectives. This push is partly due to claims in the Spellings Report (US Department of Education, 2006) and other influential publications (e.g. Arum & Roksa, 2011; Hart Research Associates, 2015) that students do not leave higher education with the knowledge and skills necessary to be successful in the workforce. Though there are many advocates for the use of performance assessments in higher education, the American Association of Colleges & Universities (AAC&U) is perhaps the most prominent voice. To facilitate the use of performance assessments in higher education, the AAC&U released the Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics in 2009. The VALUE rubrics are a set of 16 rubrics developed to assess critical learning objectives for higher education (AAC&U, 2015). In addition to being used to assess student learning objectives, AAC&U proposed

that the VALUE rubrics could be adapted for use in classrooms to facilitate formative learning, or as large-scale assessment tools to summatively evaluate students' abilities to meet collegiate learning objectives (AAC&U, n.d.).

Several years later, AAC&U and the State Higher Education Executive Officers unveiled The Multi-State Collaborative to Advance Student Learning Outcomes Assessment (MSC), a framework developed to promote the use of performance assessments in higher education. The MSC initiative was developed to assist institutions in using course-embedded performance assessments as part of a nationally-organized assessment process (AAC&U, 2017), and was partly developed in response to negative perceptions surrounding selected-response exams. The MSC is perhaps the most prominent large-scale performance assessment initiative present in today's higher education landscape. Currently, thirteen states and over 70 two- and four-year institutions participate in the MSC (AAC&U, 2017). Given the success and popularity of the MSC thus far, it is likely that the initiative will continue to scale up, expanding to additional states and/or institutions. With the expansion of the MSC will come an increase in the use of performance assessments to evaluate institutional learning objectives.

Though performance assessments are popular in today's higher education context, they are prone to psychometric challenges that hinder widespread adoption and use. Perhaps one of the largest challenges is the subjective nature of performance assessment scoring. Performance assessments involve carrying out a process or creating a product, which can often only be scored by human raters exercising judgment to determine the extent to which students met pre-specified scoring criteria. This scoring process is in contrast to selected response assessments, where a single option is often correct, resulting

in what some consider to be an objective scoring process. Performance assessments are often scored by human raters, resulting in scores that are rater-mediated and possibly a product of rater judgment in addition to, or instead of, student ability (Engelhard, 2002). The subjective nature of rater-mediated scoring raises questions about what it is that scores represent (Stiggins, 1987), and educators must provide evidence that performance assessment scores are a function of student ability, not a function of raters (AERA, APA, & NCME, 2014). Many researchers question whether raters can actually be an objective channel through which scores are produced (Guilford, 1954; Schafer, Gagné, & Lissitz, 2005). Often, evidence suggests that rater characteristics permeate performance assessment scores, resulting in decreased psychometric quality of scores, and questions regarding score utility (Cizek, 1991a).

Even the best performance assessment systems are not immune from the psychometric challenges. At James Madison University, the same is true for Ethical Reasoning in Action: The Madison Collaborative, an institution-wide ethical reasoning program. The Madison Collaborative implements several performance assessments, one of which is the focus of this dissertation research. As such, a description of the Madison Collaborative is provided as context for this study.

**Ethical Reasoning in Action: The Madison Collaborative**

In 2011, James Madison University adopted Ethical Reasoning in Action: The Madison Collaborative as its Quality Enhancement Plan for regional accreditation (James Madison University, 2013). The Madison Collaborative has proposed a specific ethical reasoning framework, through which students should engage when considering an ethical situation. The framework was built on the idea that ethical reasoning involves asking

relevant, open-ended questions that assist the decision maker in understanding ethical situations and their multi-faceted natures (Sanchez, Fulcher, Smith, Ames, & Hawk, 2017).

The framework is operationalized by eight Key Questions (8KQ): Fairness, Outcomes, Responsibilities, Character, Liberty, Empathy, Authority, and Rights. Related to each Key Question (KQ) word is a question to consider when making an ethical decision (Sanchez et al., 2017):

- Fairness: How can I act equitably and balance legitimate interests?
- Outcomes: What achieves the best short- and long-term outcomes for me and all others?
- Responsibilities: What duties and/or obligations apply?
- Character: What action best reflects who I am and the person I want to become?
- Liberty: How does respect for freedom, personal autonomy, or consent apply?
- Empathy: What would I do if I cared deeply about those involved?
- Authority: What do legitimate authorities (e.g. experts, law, my religion/god) expect of me?
- Rights: What rights (e.g. innate, legal, social) apply?

Note that, though a single question is posed for each KQ, each question is merely an example and starting point for using the KQ framework. When exhibiting facility with the 8KQs, students determine which KQs are most relevant to their given ethical situation, analyze the relevant KQs, and balance multiple questions related to each KQ to come to a decision. For example, consider a student who is integral to financially supporting his family, and he is faced with the decision to enter the workforce or attend

college after his high school graduation. He may consider the key question Outcomes in the context of his ethical decision. He may ask what the short- and long-term outcomes will be regarding his relationship with this family if he decides to go to college. Will he be able to preserve his familial relationships if he goes to college? Will he become the family outcast? He may ask what the short- and long-term repercussions will be for his family. Will his parents find venues to support themselves? Will they be able to pay their rent? He may also consider Authority. He may ask whom the authority figures are in his situation and wonder what those figures expect of him. What do his parents expect of him? What does his school counselor expect of him? Together, the KQs provide a framework for students to use as a guide when making ethical decisions. Each KQ provides a base from which students can ask additional questions and deepen their understanding of the complexities of their ethical situations.

Coordinators of the Madison Collaborative develop programs under the premise that ethical reasoning is a skill that can be learned through thoughtful and targeted interventions (Sanchez et al., 2017). Seven student learning objectives (SLOs) outline what students should be able to know, think, or do as a result of participating in Madison Collaborative interventions:

1. Students will be able to state, from memory, all eight Key Questions.
2. When given a specific decision and rationale on an ethical dilemma, students will correctly identify the Key Question most consistent with the decision and rationale.
3. Given a specific scenario, students will identify appropriate considerations for each of the Eight Key Questions.

4. For a specific ethical situation or dilemma, students will evaluate courses of action by applying (weighing and, if necessary, balancing) the considerations raised by the Key Questions.
5. Students will apply SLO 4 to their own personal, professional, and civic ethical cases.
6. Students will report that they view ethical reasoning skills as important.
7. Students will report increased confidence in their ability to use the ethical reasoning process.

To enable students' mastery of the objectives, the Madison Collaborative designs and implements intentional curricula to guide students' facilitation with ethical reasoning. A key intervention includes *It's Complicated,* a 75-minute guided discussion in which 4,500+ first-year students participate. Students are divided into small groups of less than 40 students, and trained faculty and staff facilitate a 75-minute guided discussion. Though the 75-minute program is the only formal program all students experience, students may be exposed to the 8KQs and ethical reasoning in their coursework and/or co-curricular experiences. Thus, the level of exposure to the 8KQs varies widely across students.

To assess their seven SLOs, the Madison Collaborative has collected data related to their objectives since 2012. Of particular interest in this study is the Ethical Reasoning and Writing (ER-WR) essay assessment used to assess SLO 5. The ER-WR is a constructed-response assessment on which students are asked to compose an essay describing 1) an ethical situation with which they are familiar, 2) the ethical considerations relevant to the situation, 3) their ethical reasoning process, and 4) the

decision they made (see Appendix A for ER-WR instructions and prompt). Trained raters score the ER-WR essays using the ER-WR rubric (see Appendix B). The ER-WR rubric was developed by ethical reasoning and assessment experts, and each of the five rubric elements was designed intentionally to cover the steps through which students should progress when faced with an ethical situation.

For all ER-WR rubric elements, scores range from 0 – 4, with a score of zero considered "insufficient" and typically representing no demonstration of a skill in a student's essay. On the other hand, a score of four is considered "extraordinary," typically representing a deep understanding of all KQs and seamless integration of KQs in the ethical decision-making process. Per the JMU Strategic Plan, students should achieve, on average, a score of two or better on the ER-WR rubric by 2020 (JMU Office of Institutional Research, 2017). Though the Strategic Plan states that students should achieve, on average across all elements, a score of two, ideally students will meet a score of at least two on each rubric element. If students achieve a score of two on each rubric element, it suggests that students can 1) explicitly describe decision options related to a personal ethical situation (Element A); 2) reference four KQs (Element B); 3) provide a rationale for the applicability or inapplicability of four KQs to their ethical situation (Element C); mostly accurately apply at least three KQs to their ethical situation (Element D); and weigh the KQs and other relevant factors to come to a decision that can conceivably be derived based on the weighing and balancing of KQs and other relevant factors (Element E).

Unfortunately, on average, students historically have not met the university-determined benchmark. In effect, ethical reasoning has been the focus of various small-

scale curricular and pedagogical interventions designed to increase students' ethical reasoning skills (e.g. Good, 2015; Smith, 2017). Fortunately, Good (2015) and Smith (2017) found that ethical reasoning can in fact be learned, and the new challenge is how the institution may scale the curricular interventions to affect more students. Both of the aforementioned studies used ER-WR scores as metrics to evaluate the extent to which students' ethical reasoning improved as a result of targeted ethical reasoning interventions.

Clearly there is institutional investment related to students' ethical reasoning skills, and ER-WR scores are the foundation upon which many inferences are made regarding students' ethical reasoning abilities. Thus, to draw accurate inferences from ER-WR scores, it is imperative that ER-WR scores have solid psychometric evidence to support their uses and interpretations. In the case of the ER-WR, anecdotal and empirical evidence suggest there may be concern regarding the meaning of students' ER-WR scores. For example, raters anecdotally report that they find it difficult to distinguish between the KQs of Liberty and Rights, and therefore find it challenging to assign scores when those KQs are present in students' essays. If raters cannot distinguish between Liberty and Rights, they may assume students are analyzing the same KQ, rather than recognizing that students are analyzing two separate KQs. In this situation, students' scores may be lower than they should be.

Moreover, raters anecdotally report that it is challenging for them to distinguish between some scoring criteria. Raters report lack of clarity regarding what it looks like for a student to analyze a KQ, and how that differs from providing a rationale for the KQ's applicability or inapplicability to the ethical situation. If raters cannot distinguish

between providing a rationale for a KQ and analyzing a KQ, students' scores may or may not be accurate representations of their abilities.

Empirically, generalizability theory analyses reveal that ratings have had lower than desirable inter-reliability over the past several years (Bashkov, Smith, Fulcher, & Sanchez, 2014; Holzman, Ames, & Pyburn, 2017; Smith, Fulcher, & Pyburn, 2015; Smith, Pyburn, & Ames, 2016). G-coefficients have ranged from 0.69 to 0.75 for first-year student scores, and 0.57 to 0.66 for second-year student scores. These estimates suggest that there is considerable error variability in scores, particularly for second-year students. In previous years, variability due to raters ranged from 7%-9%, and variability due to differences in how raters rate the same student (i.e. a rater by student interaction) accounted for approximately 15% of variability in second-year students' scores. In sum, previous g-theory results suggest that ER-WR scores contain large proportions of error variability, seemingly due to rater differences. As such, there is concern regarding the extent to which scores represent students' ethical reasoning abilities. Though there could be many reasons for less than adequate reliability, a possible explanation could be that raters do not use the rubric in similar ways. For example, some raters may apply the scoring criteria more stringently than others, resulting in differential severity/leniency across raters. Differences in rater severity contribute to rater error variability, which may decrease reliability. Given raters' reported difficulty distinguishing between some of the ER-WR rubric elements, students' scores may be unnecessarily similar across elements, potentially resulting in low reliability due to a restriction of variability among scores.

Essentially, raters may interpret the ER-WR rubric differently, resulting in systematic differences in scores due to raters. These systematic differences in scores due

to raters are known as rater effects (Myford & Wolfe, 2003). When rater effects are present, scores do not solely represent students' ethical reasoning abilities, but instead represent a mix of students' ethical reasoning abilities and rater characteristics. Though the presence of rater effects in ER-WR scores has been explored via generalizability theory analyses, rater effects analyses that allow for an evaluation of individual raters have not been explored. Given that ER-WR scores are used to make institution-level inferences regarding students' ethical reasoning abilities, it is warranted to further investigate ER-WR scores for rater effects.

**Study Purpose & Research Questions**

The purpose of the current study is to evaluate ER-WR scores for rater effects. Of specific interest is the evaluation of rater leniency/severity and restriction of range. Additionally, as raters' knowledge of the 8KQs may influence their scores, an additional purpose of the study is to identify whether there are systematic relationships between raters' knowledge of the 8KQs and rater effects.

Evaluating ER-WR scores for rater effects will be useful for the Madison Collaborative moving forward. If scores are not influenced by rater effects, this study provides further validity evidence to support that ER-WR scores represent students' ethical reasoning abilities. If scores are influenced by rater effects, this study provides further information regarding rater behaviors and the relationship between rater effects and rater 8KQ knowledge. This information is useful for the Madison Collaborative, as it may have implications for the interpretations of ER-WR scores, as well as rater training and selection.

In this study, the following research questions were addressed:

1) Are there statistically significant differences in rater leniency/severity, suggesting a group-level leniency/severity rater effect?
2) Are there statistically significant rater leniency/severity and ER-WR rubric element interaction effects?
3) Is there a lack of distinguishability between score levels, suggesting a restriction of range effect? Is this restriction of range suggestive of a central tendency effect?
4) Is there a relationship between raters' knowledge of the 8KQs and rater leniency/severity?

# Chapter 2: Literature Review

Performance assessments are popular in higher education assessment (Kuh et al., 2015). As mentioned, the American Association of Colleges & Universities (AAC&U) has promoted several performance-based assessment systems, including the Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics and the Multi-State Collaborative (MSC). Moreover, many institutions have chosen to implement their own home-grown performance assessments, independently of nationally-organized higher education performance assessment systems. A primary impetus for the popularity of performance assessments is the claim that performance assessments allow for better evaluation of higher order thinking and learning, in comparison to selected-response assessments. However, there are many challenges related to performance assessments. Specifically, because performance assessments often require human raters to score students' products, many psychometric challenges present for performance assessment scores. One psychometric challenge is that of rater effects, or systematic differences in how raters rate students' products (Myford & Wolfe, 2003). Rater effects have potentially grave implications for the inferences made from performance assessment scores. The purpose of this literature review is to describe the advantages and disadvantages of performance assessments, leading up to a discussion of rater effects as a concern for performance assessment scores.

## Performance Assessments

Performance assessments are comprised of two main components: 1) the performance task and 2) the scoring of that task (Khattri, Reeve, & Kane, 1998). With performance assessments, the performance task requires students to construct a product

and/or carry out a process. The product or process is then evaluated via a rater, either after the product has been completed, or while the process is being carried out by the student (Johnson et al., 2009). For example, an art major may be required to assemble a portfolio as a capstone project for the major. The portfolio may be evaluated by the faculty based on demonstration of color and texture, creative idea, or improvement over time. Or, a chemistry 101 student may be asked to perform a titration and be evaluated by faculty based on ability to adequately perform the titration process. The common thread for both examples is that the student is *creating* a product or *performing* a task that is scored via trained raters.

Performance assessments are often referred to as constructed-response assessments, alternative assessments, or authentic assessments. Performance assessments are described as *constructed-response* assessments due to the fact that students are required to construct a product or construct a process for carrying out a task. This is in contrast to selected-response assessments, on which students are asked to select the best answer from a list of possible answers (Downing, 2006). The constructed-response nature of performance assessments is often thought to be *alternative* to selected-response assessments (Wiley & Haertel, 1996). That is, students must demonstrate their knowledge and abilities through action in performance assessments, but must demonstrate their knowledge through selection in selected-response assessments, providing the basis for the "alternative" language surrounding performance assessments. Because performance assessments require engagement in a process or completion of a product, they are often thought to have better fidelity to real-life situations than selected-response assessments, on which students must simply select correct answers. Critics of selected-response

assessments argue that selected-response assessments are decontextualized and do not represent true-to-life scenarios. The increased fidelity to real-world situations results in performance assessments being coined as *authentic* assessments (Linn, Baker, & Dunbar, 1991; Stecher, 2014; Wiggins, 1991).

As can be evidenced by the language surrounding performance assessments, a tension exists between performance assessments and selected-response assessments (Cizek, 1991a, 1991b; Wiggins, 1991, 1993). Tensions between these two assessment methodologies primarily arise from three arguments: 1) types of knowledge and thinking each type of assessment is able to assess, 2) logistical concerns and resources necessary to implement each type of assessment, and 3) psychometric properties of scores from each type of assessment.

**Types of knowledge best assessed by performance assessments.** Selected-response assessments provide a legitimate means of measuring knowledge (Downing, 2006; Haladyna, 2004); however, they may fall short when measuring higher-order knowledge, skills, and/or abilities. Given that students are required to create a product or engage in a process during a performance assessment, performance assessments are perceived by some as better than selected-response assessments in eliciting students' higher-order thinking (Lane & Stone, 2006; Wiggins, 1991). Performance assessment tasks tend to be complex and integrate real-life context into the assessment, requiring students to synthesize and implement knowledge to demonstrate proficiency in a realistic situation (Linn et al., 1991). Moreover, students may *know* something, but being able to apply and *perform* knowledge is an additional skill that is challenging to measure through a selected-response format.

In addition, performance assessments allow for a scaffolded system of learning (Gronlund, 2003), thus representing how students learn and use knowledge to develop higher-order thinking and skills (Darling-Hammond, 2014). Moreover, educators recognize they cannot teach all knowledge necessary for the workforce, so must instead teach students to develop higher-order thinking abilities and skills, as these skills will assist students in their success after K-12 and higher education (Lenz, Wells, & Kingston, 1991). Performance assessments are thought to simulate how students learn, synthesize, and apply knowledge and skills, thus mirroring the process in which students must engage to be successful in the workforce. The "best" performance assessment systems are those in which students learn from the assessment process, possibly by receiving feedback and/or allowed the opportunity to revise and re-submit their work (Gronlund, 2003; Welch, 2006). This iterative process of completing the assessment, receiving feedback, and having the opportunity to revise the assessment facilitates students' learning and higher-order thinking capabilities (Wiggins, 1998).

Many educators advocate for performance assessments on the basis that they are more direct measures of students' higher-level thinking abilities compared to selected-response assessments (Lane & Stone, 2006; Resnick & Resnick, 1996). However, it is also important to consider that a performance assessment in its own right will not necessarily evoke desired higher-order thinking, nor will the assessment inherently align with real-world context (Linn et al., 1991). A tremendous amount of time, thought, and effort must be put into any assessment to ensure that it evokes the necessary skills and knowledge (Schmeiser & Welch, 2006), and, if poorly developed, performance assessments will not elicit higher-order thinking. Thus, just like any assessment,

performance assessments require careful development of the assessment task. Moreover, performance assessments require vast resources to administer and score.

**Logistical and resource concerns associated with performance assessments.** It is generally accepted that performance assessments require more resources than selected-response assessments (Downing, 2006; Gronlund, 2003; Linn et al., 1991; Madaus & Kellaghan, 1993). Similar to selected-response assessment, performance assessment development requires highly skilled task writers, piloting of tasks, revising tasks, and preliminary data collection for validity evidence (Welch, 2006). However, because performance assessments are subjectively scored, they also require the development of a scoring guide. The scoring guide most often takes the form of a checklist or rubric (Johnson et al., 2009) and is an integral component for ensuring scores are meaningful and useful representations of students' abilities (AERA, APA & NCME, 2014; Khattri et al., 1998; Stiggins, 1987).

After the development of the assessment prompt and the scoring guide, students must complete the assessment. After students complete the assessment, trained raters must rate the products. It is generally recommended that at least two raters score each product (Johnson et al., 2009). Thus, a considerable amount of time and resources are dedicated to the scoring of performance assessment products after students complete the assessment.

The logistical concerns related to performance assessments quickly compound if educators desire a broad assessment of student knowledge and skills. Compared to selected-response assessments, students are not able to complete as many performance-based tasks in the same amount of time it takes them to complete selected-response tasks

(Downing, 2006; Gronlund, 2003; Linn et al., 1991). Thus, if educators want to evaluate students broadly on a construct, students must complete several performance assessments and devote an immense amount of time to the testing process. Typically, educators do not have the amount of time available that is necessary to broadly cover a construct with a performance assessment. Thus, performance assessments provide a logistical challenge if the purpose of the assessment is to obtain a broad depiction of student knowledge and abilities.

Though estimates of the cost of performance assessments are variable (Picus, Adamson, Montague, & Owens, 2010), substantial costs are associated with the development, administration, and scoring of performance assessments (Hardy, 1995). Proponents of performance assessments argue that the sustained costs are worthwhile, particularly if the data obtained are accurate measures of higher-order learning and represent what students are capable of in a real-world context (Hardy, 1996; Picus et al., 2010; Wiggins, 1993). However, educators have a responsibility to consider the costs associated with performance assessments (Cizek, 1991b; Topol, Olson, & Roeber, 2010). This responsibility is particularly important considering that performance assessment scores may suffer from poor psychometric quality (Cizek, 1991b; Downing, 2006), in effect raising concerns about what scores represent (Bejar, 2012).

**Psychometric properties and trustworthiness of scores from performance assessments.** Psychometric concerns surrounding performance assessment scores may stem from several sources, and there are often more concerns about the psychometric properties of scores from performance assessments than selected-response assessments. Concerns particularly arise due to the more challenging nature of evaluating performance

assessment scores than selected-response scores. For example, it is often more challenging to run and interpret polytomous item response theory (IRT) models used for performance assessment data than it is to run and interpret dichotomous IRT models often used for selected-response assessment data. Moreover, it may be more challenging to claim that performance assessment scores adequately represent the construct of interest compared to selected-response assessment scores (Brennan, 2001).

To make valid inferences regarding students' abilities on the construct of interest, the assessment must cover the content of the construct with adequate breadth (AERA, APA & NCME, 2014). That is, when considering the meaning of scores, educators must consider the extent to which evidence supports the generalization of the scores to the construct of interest (Brennan, 2001; Haertel, 1999). As discussed above, it is challenging to achieve breadth of a construct with performance assessments. Though performance assessments may yield important information regarding students' depth of understanding of a construct, the lack of breadth provides a limitation to the generalizability of performance assessment scores to a construct (Brennan, 2001; Messick, 1996).

In large-scale testing situations, educators may desire to develop several performance assessment tasks in order to maintain test security (Picus et al., 2010). However, as with all assessments, information regarding task comparability must be provided to ensure that students' scores are not a function of which task they received (AERA, APA, & NCME, 2014). Comparability across performance assessment tasks may be challenging to achieve, and differences in performance assessment tasks may contribute substantially to differences between students' scores (Shavelson, Baxter, & Gao, 1993; Hathcoat, Penn, Barnes, & Comer, 2016). To account for test form

differences in selected-response assessment situations, equating is an often go-to psychometric fix to remove variability related to minor differences between test items (Bandalos, 2018; Wendler & Walker, 2006). However, the same equating procedures are challenging with performance assessment tasks (Lane & Stone, 2006; Muraki, Hombo, & Lee, 2000), thereby introducing an additional psychometric challenge into the performance assessment process.

An additional consideration regarding the trustworthiness of scores is the manner of scoring. Performance assessments do not typically have a clear correct or incorrect response. Rather, scoring is a subjective, rater-mediated process performed by human raters or computer algorithms (Engelhard, 2002; Johnson et al., 2009), and the subjective nature of the scoring process provides an additional avenue through which error may be introduced into scores (Linn, 1993). For performance assessments, evidence must be presented to demonstrate that scores primarily reflect students' abilities, not the rater who rated the student work (AERA, APA, & NCME, 2014). A sound development process for the scoring guide is a critical first step in providing evidence of score interpretations (Welch, 2006). However, even with a well-developed scoring guide and rater training, raters tend to interpret and use scoring guides differently from one another (Barkaoui, 2007; DeRemer, 1998; Holzman, 2016; Huot, 1990; Rezaei & Lovorn, 2010). The extent to which raters differ in their use of the scoring guide limits the validity of scores as representations of student ability on the construct of interest. Thus, a critical question often asked regarding performance assessment scores is if scores actually represent student ability, or if they represent some conglomerate of ability and rater characteristics.

This question about the validity of the interpretations of scores is exacerbated by the confusing nature of reliability for performance assessment scores, as reliability may be conceptualized in several ways for performance assessment scores (Stemler, 2004). Specifically, reliability is often operationalized as consensus, consistency, or agreement between raters. In addition to different types of reliability, information gleaned from these types of reliability may contradict themselves. For example, raters may exhibit high consistency across students, but poor agreement may be observed between raters (Eckes, 2015; Stemler, 2004). Such an outcome may occur if one rater is relatively severe while another rater is relatively lenient. Students would be rank-ordered similarly across raters, resulting in high consistency. However, there would be low agreement between raters, as their scores are not perfect matches of one another. This contradictory outcome may be confusing for educators and researchers.

To exacerbate the issue, many researchers do not explicitly state a rationale for the type of inter-rater reliability or agreement they use. These seeming contradictions between information gleaned from different reliability indices, coupled with the lack of explicit reference to types of reliability creates confusion for stakeholders and introduces further questions regarding the trustworthiness of performance assessment scores. For example, in a review of performance assessment studies, Jonsson and Svingby (2007) found that few studies achieved interrater agreement of 0.70 or higher. Such findings might suggest that performance assessment scores have limited reliability, thus creating distrust around the meaning of scores. However, interrater agreement is a rather stringent form of reliability, and for many low-stakes educational assessments, the stringency of interrater agreement may not be necessary. Moreover, some interrater agreement indices

(e.g. Kappa) are heavily influenced by the prevalence of scores across the score levels (i.e. many scores piled up in some score levels), at times making it appear as though inter-rater agreement is particularly low, when it may in fact be acceptable (Gwet, 2014). In short, reliability for performance assessment scores is a nuanced topic and ambiguity in research may lead to confusion and additional psychometric concerns that may or may not be founded.

The additional resources and perceived psychometric challenges of performance assessments should not necessarily prompt educators to forgo performance assessments in favor of selected-response assessments. Rather, the choice between selected-response or performance assessments should stem from the purpose of the assessment; the content, cognitive level, breadth, and depth to be assessed; and logistical considerations (Lane & Stone, 2006; Schmeiser & Welch, 2006). Moreover, with a sound development process, performance assessments may be effectively used to gather information regarding student knowledge and abilities. Because the purpose of this study was to evaluate performance assessment scores for rater effects, and various rater effects may arise as a function of rubric design, rubrics are discussed in more detail.

**Rubrics**

Rubrics are the most common scoring guide used to score performance assessments (Saal, Downey, & Lahey, 1980), and rubrics are imperative for achieving adequate psychometric properties of scores (Welch, 2006). There are two traditional types of rubrics: holistic and analytic. As the name implies, holistic rubrics are used to evaluate a performance assessment product or process *holistically.* That is, the construct is not assessed with separate elements, but is instead assessed with only one element,

resulting in a single score that is representative of students' holistic performance on the task (Gronlund, 2003; Huot, 1990; Lane, 2014). In contrast, analytic rubrics allow various elements of the construct to be evaluated individually (Moskal, 2000; Welch, 2006), rather than synergistically as one element with a holistic rubric. Thus, multiple scores will be generated for a single performance assessment product or process when an analytic rubric is used, and only one score will be generated when a holistic rubric is used. Neither type of rubric is better nor worse than the other; the type of rubric depends on the theory underlying the construct (Wiggins, 1998) and the type of information desired from the assessment (Lane & Stone, 2006).

However, note that it is important that researchers and educators adequately consider whether their rubric should be holistic or analytic. If analytic rubric elements are too similar to one another, raters may be unable to differentiate between them, resulting in similar scores across rubric elements (DeCotiis, 1977; Johnson et al., 2009). Conversely, if a holistic rubric is used when there are actually several different elements of a construct, raters may be unsure how to prioritize each dimension when deriving a score. Or, a product or process may encompass some features of higher scores and some features of lower scores for different dimensions, resulting in confusion about how to appropriately provide a score to the product or process (Barkaoui, 2007). Thus, if developing a rubric, the choice between a holistic or analytic structure is important and may influence the psychometric quality of ratings.

Scoring criteria may also influence the manner in which raters rate, potentially influencing the psychometric properties of scores. If the rubric is holistic, a single scoring criterion that encompasses all relevant skills should be developed for each score level. If

the rubric is analytic, scoring criteria should be developed for each score level on each element. Regardless of whether a rubric is holistic or analytic, a rubric should be designed in such a way that raters are able to use the scoring criteria to differentiate students of varying abilities (Johnson et al., 2009). Rubrics make explicit the notion that there is a continuum of ability underlying most skills (Wiggins, 1998), and the continuum of ability should be clearly articulated in the scoring criteria. When using the scoring criteria, raters should be able to accurately place students along the ability continuum and separate students based on their performance on the task.

To best facilitate use of the rubric and consistent scoring across raters, scoring criteria must 1) clearly define the qualities at each score level, 2) build upon the previous score, and 3) be consistent in language across the score levels (Tierney & Simon, 2004). Scoring criteria will build in intensity, quality, or quantity across the score range; however, new criteria should not be introduced at different score levels within the same dimension (Wiggins, 1998). Moreover, if possible, criteria should be described descriptively, rather than quantifying the dimension with judgments such as “a lot” or “some” (Moskal, 2000). Descriptors such as “a lot” or “some” require raters to exercise judgment, and the meaning of “a lot” or “some” may vary across raters, introducing additional subjectivity into the rating process. Instead, rating criteria should describe “a lot” or “some,” perhaps numerically. Clear descriptions of the performance criteria at various levels assists raters in accurately differentiating students (Moskal & Leydens, 2000). In sum, across the scoring levels, each of the scoring criteria should evaluate the same content within the dimension, the language used to differentiate between scores should be clear and consistent, and the score criteria should logically build across the

scoring levels. Unfortunately, consistency of rubric criteria is not often discussed in rubric development literature, resulting in a lack of understanding of the necessity for consistent scoring criteria in rubric development (Tierney & Simon, 2004).

Closely tied to the scoring criteria is the score range and number of score levels. There is no explicit rule for how many score levels a rubric should have. However, it is generally accepted that rubrics should include enough scoring levels to clearly differentiate between students' performance on each element, but not so many or so few scoring levels that the distinctions between levels is indistinguishable or muddied (Lane & Stone, 2006). For example, a rubric with many score levels may create confusion for raters and result in an inability to differentiate between the criteria at the score levels, ultimately resulting in raters using the same middle score levels (Landy & Farr, 1980).

In large-scale assessments, the primary intended purpose of rubrics is to guide raters through the scoring process (Lane & Stone, 2006). By making explicit the qualities that are most valued in the task and specifying what various levels of achievement look like for each quality, rubrics aid in systematizing the way in which raters score performance assessment products and processes (Johnson et al., 2009; Tierney & Simon, 2004). Rubrics provide a scoring structure for raters, thus making scoring less subjective and, in effect, increasing credibility for performance assessment scores. The extent to which the scoring process can be shown to be the same across raters lends support for the claim that scores represent student ability rather than rater characteristics (Stiggins, 1987).

Despite high-quality rubrics, raters may still have tendencies (e.g. general harshness or leniency) that influence their ratings (Gronlund, 2003). Ideally, differences

between raters will be negligible (Eckes, 2009). However, differences between raters often are not negligible, resulting in repercussions on the psychometric quality of students' scores. To the extent that rater differences are not negligible, construct-irrelevant rater variance is introduced into scores. These differences may take different forms and are referred to as rater effects (Myford & Wolfe, 2003).

**Rater Effects**

Though educational researchers strive to create an objective scoring process through rater training and well-developed rubrics, ratings remain deeply rooted in rater judgment (Eckes, 2009; Myford & Wolfe, 2003). Performance assessment ratings have been referred to as "rater-mediated," as they represent raters' perceptions of students work, raters' interpretations of the rubric, and the raters' analysis of how the student work and the rubric align (Engelhard, 2002). Raters' perceptions of the rubric and how the rubric should be applied to student work may or may not align with the intended interpretations and uses of the rubric.

To improve alignment between raters' interpretations of the rubric and the intended interpretations of the rubric, rater training is often implemented. However, even with rater training, raters' interpretations of the rubric may not align with one another, resulting in systematic differences in students' scores across raters. Systematic errors in raters' scores that reflect raters' personal characteristics and/or personal interpretations of the rubric are known as rater effects (Bond & Fox, 2015; Eckes, 2009; Myford & Wolfe, 2003; Scullen, Mount, & Goff, 2000). The most commonly discussed rater effects are leniency/severity, halo, central tendency, and restriction of range (Myford & Wolfe, 2003).

**Leniency/severity.** As discussed, raters ideally interpret a rubric in the same way. Specifically, all raters ideally adopt the same scoring criteria and apply these scoring criteria 1) consistently across all students, and 2) in the manner intended by rubric developers. However, raters may adopt the same scoring criteria, yet vary in how stringently they apply the scoring criteria (Wolfe, 2004). Leniency and severity are characterized by raters consistently assigning high or low scores, respectively, across examinees (Eckes, 2009, 2015; Engelhard, 1992; Saal et al., 1980). Raters are considered severe if they consistently assign low scores across all examinees, and raters are considered lenient if they consistently assign high scores across all examinees. Said differently, severe raters are those whose average ratings are lower than the average ratings assigned by all raters, and lenient raters are those whose average ratings are higher than the average ratings assigned by all raters (Bond & Fox, 2015; Eckes, 2015; Wolfe, 2004).

Given that scores are thought to be a proxy for student ability, consistently severe or lenient scores are problematic, as students' abilities are either under- or over-estimated. Ideally, all raters will be of similar average rating severity (Myford & Wolfe, 2004). Moreover, raters are often assumed to be of similar rating severity in most research (Lunz, Wright, & Linacre, 1990). However, raters are often found to vary drastically from one another in their severity (Eckes, 2005; Han, 2014; Lunz et al., 1990).

Moreover, as discussed above, raters ideally rate in a manner that is consistent with the intended interpretations and uses of the rubric. Expert ratings are often used to represent scores that reflect the intended interpretations and uses of the rubric. Expert raters are most often content experts who are highly familiar with the content of the

assessment as well as the rubric (Johnson et al., 2009). Thus, rater severity or leniency may be gauged by comparing rater scores back to expert scores. Raters often differ from expert raters in their leniency and severity (Engelhard, 1994). Differential severity or leniency across raters may have dire consequences, particularly if high-stakes decisions are made from scores. In fact, differential severity or leniency across raters may result in inaccurate placement decisions for examinees (Lunz et al., 1990; Wu & Tan, 2016; Yan, 2014).

Raters' leniency or severity may also change over a rating period. Much of the literature suggests raters tend to become more severe over time, particularly across rating periods of several days or more (Congdon & McQueen, 2000; Leckie & Baird, 2011; Pinot de Moira, Massey, Baird, & Morrissy, 2002). However, in a study evaluating rater effects in AP English Literature and Composition essays, Wolfe, Myford, Engelhard, and Manalo (2007) found that only 5% of raters become more severe over the rating period, while 16% of raters became more lenient over the rating period. Thus, it appears that raters' leniency and severity may change over time, and the direction of the change may not always be predictable. As such, the design of the rating session is of concern in large-scale assessment situations where raters are expected to rate over multiple days (Congdon & McQueen, 2000). If resources allow, educators or researchers may opt for additional raters in order to shorten the rating period.

Raters' leniency and severity may also vary across rubric dimensions. That is, raters may rate more severely on some rubric dimensions compared to other rubric dimensions. In a study related to the writing assessment in the Test of German as a Foreign Language, Eckes (2005) found that more than one-third of raters exhibited

differential severity across rubric elements. These results suggest that raters may be inconsistent in the stringency that they apply across various rubric elements, implying potential interaction effects between raters' leniency and severity and rubric elements. An interaction between rater leniency/severity and rubric element may also be referred to as differential rater functioning or bias (Eckes, 2015). Such an effect may be particularly problematic in compensatory models where students are awarded differential credit by rubric element.

Additionally, raters' leniency and severity may not be constant across scoring levels. In a study related to an Oral English Proficiency Test, Yan (2014) found that raters differed in their severity or leniency depending on score level. Specifically, raters were more similar to one another for tests that scored on the passing side of the score levels than for tests that scored on the failing side of the score levels. As such, raters may be unable to consistently determine the meaning of scoring criteria across score levels. This effect may be particularly problematic in instances where there is a passing score that students must meet in order to be awarded placement into a program, awarded certification, awarded scholarship money, etc. Though rater leniency and severity is perhaps the most heavily researched rater effect, the halo effect has also been heavily researched.

**Halo.** When rating student work, raters develop an initial impression of the product, and they then have to balance this impression with the proposed scoring criteria defined in the rubric (Lumley, 2002). Ideally, raters forego their initial impressions and rate the product based on the criteria presented in the rubric. However, raters may struggle to objectively consider the product in the context of the scoring criteria.

Moreover, they may not recognize the extent to which their initial impression of the product influences the scores they assign (Nisbett & Wilson, 1977). Inability to ignore overall, initial impressions of students' products may manifest as a halo effect (Fisicaro & Lance, 1990; Humphry & Heldsinger, 2014; Eckes, 2015; Myford & Wolfe, 2003; Thorndike, 1920).

The halo effect is characterized by the phenomenon in which raters cannot differentiate between distinct rubric elements, resulting in highly correlated scores across elements for a single product (Borman, 1975; Saal et al., 1980). As alluded to above, halo most often occurs when raters perceive a general, global impression of the product, and this global evaluation hinders raters' abilities to evaluate distinct rubric elements (Thorndike, 1920). When a halo effect occurs, it creates an inaccurate dependency among distinct rubric elements, resulting in a scoring schema that more closely resembles a holistic schema, rather than an analytic scoring schema (Engelhard, 1994).

The presence of a halo effect may be readily apparent by a quick visual examination of ratings. For example, suppose one rater provided rubric element ratings of 2, 2, 3, 2 for an essay. Suppose another rater provided rubric element ratings of 2, 4, 1, 4 for the same essay. Because of the first rater's similarity in ratings across elements, the first rater's scores are more likely to have been influenced by a halo effect than the second rater's scores. It is important to note that, though similar scores across rubric elements is an indicator that a halo effect might be present, it does not absolutely indicate that a halo effect is present (Murphy & Cleveland, 1991; Solomonson & Lance, 1997). It could be the case that students' abilities are actually similar across elements, so similar scores across elements is warranted and accurate. Consider the previous example. If the

student's abilities were similar across all rubric elements, then the first rater's scores may actually be more accurate than the second rater's scores. However, if the student's abilities were not similar across the four rubric elements, then the first rater's scores are likely inaccurate and suggest the rater may be exhibiting a halo effect.

As mentioned, a halo effect often occurs due to a global impression of the product that clouds raters' abilities to rate each rubric element independently. However, a halo effect could also occur if the rubric scoring criteria are not clearly differentiable (Nisbett & Wilson, 1977). Recall that an analytic rubric should be designed in such a way that the important features of a construct are defined through different rubric elements. However, there should not be so many rubric elements that raters cannot distinguish between them. When raters cannot distinguish between rubric elements due to substantial content overlap, similar ratings across the elements will be observed.

In a similar vein, the number of scoring levels for rubric elements may influence halo effects. Specifically, Humphry and Heldsinger (2014) hypothesized that restraining the number of scoring levels to be consistent across rubric elements may induce a halo effect, as constraining elements to be of the same number of scoring levels may induce a conceptual similarity across elements that may not actually be present. When raters used a rubric with varying numbers of scoring levels across the rubric elements, a halo effect was minimized. Thus, it could be the case that the number of scoring levels influences a halo effect, and researchers may minimize halo effects by allowing rubric elements to differ in their number of scoring levels. Additionally, halo effects may be minimized by asking raters to rate all student products on one element only, before moving on to rate subsequent elements (Myford & Wolfe, 2003).

A high-quality rubric development process is important for guarding against a halo effect. In addition to inability to distinguish between rubric dimensions, raters may be unable to distinguish between scoring levels, resulting in other rater effects, such as the central tendency rater effect.

**Central Tendency.** In normative assessment situations, the goal is to separate students along a continuum of ability (Bandalos, 2018; Crocker & Algina, 1986). Thus, in normative assessment situations, raters ideally use the entire score range when assigning ratings to student work. However, some raters may feel uncomfortable or averse to assigning extreme scores, so will exhibit a tendency to assign scores on the mid-point of the score range (DeCotiis, 1977; Long & Pang, 2015). This tendency is known as the central tendency rater effect (Saal et al., 1980). Central tendency effects are often prominent within rating sessions where raters are monitored and receive feedback on their ratings throughout the rating process, as raters may be less likely to provide low or high scores if they know they will receive feedback or be questioned for providing extreme scores (Myford & Wolfe, 2004; Wolfe et al., 2007).

A central tendency effect may also present if raters are unable to differentiate between the scoring criteria across score levels (Myford & Wolfe, 2004). For example, if scoring criteria are unclear, or raters cannot recognize how the scoring criteria are different across score levels, they may tend to assign ratings around the mid-point of the score range. Similar to the halo effect, a central tendency effect results in limited score variability. Note that a halo effect is closely related to the quality of individual products and results in limited score variability across rubric dimensions for individual students. A central tendency effect is a rating characteristic across all student products and may result

in limited score variability within or across students. Thus, a halo effect often results due to quality of student work, whereas a central tendency effect often results due to rater characteristics and/or unclear scoring criteria.

It appears that the central tendency effect is pervasive in performance assessment ratings. In an evaluation of essay scoring, Leckie and Baird (2011) found that, on average, most raters in their sample succumbed to the central tendency effect. Consequently, raters tended to over-rate low quality essays and under-rate high-quality essays. Engelhard (1994) found similar findings in an evaluation of essay scores. In his study, nearly 80% of ratings comprised scores from the middle two scoring levels.

Similar to the halo effect, it is important to note that an influx of scores at the mid-point of the rating scale does not necessarily indicate a central tendency rater effect, as it could be the case that students are actually of moderate ability. To disentangle whether scores are restricted to the mid-point due to student ability or a central tendency effect, researchers may evaluate the variability of ratings across students on each dimension for a single rater. For example, consider a rater who rated 20 essays on five rubric elements. Researchers may consider averaging all scores on each element across all 20 student essays and computing a standard deviation around the average for each rubric element (Saal et al., 1980). If the average is near the mid-point of the scoring levels and the standard deviation is small, it suggests that scores are clustered around the mid-point of the scoring levels and a central tendency effect may be present. Though central tendency is represented by a clustering of scores around the midpoint of the scoring levels, scores may cluster at any part of the rating scale, indicating a restriction of range effect.

**Restriction of Range.** Central tendency and restriction of range are sometimes discussed together as a single rater effect (e.g. Wolfe & Chiu, 1997). However, though central tendency and restriction of range are related, they are not necessarily the same, and may present through different patterns in the scores (Saal et al., 1980). Central tendency is a type of restriction of range in that it represents a restriction to the middle score levels (Myford & Wolfe, 2003). However, restriction of range can occur at any score level. Thus, central tendency is a type of restriction of range, but restriction of range does not necessarily imply a central tendency effect. Moreover, all rater effects previously discussed can result in a restriction of range. For example, if raters are severe, their ratings will predominately be restricted to the low end of the score levels. If raters are lenient, their ratings will predominately be restricted to the upper end of the score levels. If raters demonstrate a halo effect, similar scores will be assigned across rubric dimensions, resulting in a restriction of range at any score level. Considering any rater effects may manifest more broadly as a restriction of range effect, restriction of range is of utmost importance when evaluating ratings (Engelhard, 1994).

As with the central tendency effect, restriction of range may occur due to rater beliefs about the score levels. For example, the lowest score on a rubric is often the absence of a skill. Some raters may philosophically believe that student work is never completely devoid of a skill and consequently refrain from assigning scores at the low end of the scoring levels. The highest score on a rubric is often represented by the most exemplary demonstration of a skill. Some raters may philosophically believe that student work can always be improved, in effect refraining from assigning scores at the high end of the scoring levels. In a study of creativity assessment, Long and Pang (2015) found

that raters tended to modify the scoring levels based on their beliefs about creativity, often resulting in a restriction of range. For example, to justify a lack of scores at the low end of the scoring levels, one rater noted that "everybody possesses creativity and nobody's response is not creative" (p. 21). This rater's belief about how creativity manifests within students led to a restriction of scores to the middle to upper scoring levels.

Though all raters may succumb to rater effects, raters' backgrounds in particular may shape the manner in which they interpret student products as well as the scoring criteria, in effect inducing rater effects.

**Rater Effects and Rater Background**

Despite the fact that there is often an intended interpretation of a rubric and an intended manner in which the scoring criteria will be applied to student work, raters do not always follow the intended interpretations and applications of rubrics. Often, rubric scoring criteria and raters' backgrounds synthesize to form their own scoring schemas that they ultimately use to score the products (Bejar, 2012; Eckes, 2008; Wolfe, Kao, & Ranney, 1998). Backgrounds may be influenced by age, experience with rating, proficiency with rating, experience with the content, etc. Using their personal scoring schema, raters derive an intuitive interpretation of the product and then often use the provided scoring criteria to justify their scores (Baker, 2012; Lumley, 2002).

The extent to which raters develop their own scoring schemas may differ based on rater background. For example, in a study of raters ranging from 29 to 70 years in age, older raters perceived scoring criteria as generally less important than younger raters (Eckes, 2008). Such a finding suggests that older raters may deviate from the stated

scoring criteria more than younger raters, potentially resulting in quite different meanings of scores from older raters compared to younger raters. In the same study, older, more experienced raters placed differential emphasis on various scoring criteria compared to younger, less experienced raters who placed similar emphasis on all scoring criteria (Eckes, 2008). That is, older, more experienced raters may develop their own interpretations about which scoring criteria are most important, placing additional emphasis on certain scoring criteria when rating student work. This is in contrast to younger raters who may consider all scoring criteria to be equally important when rating. Again, such a finding suggests that scores from older raters may have quite different meanings than scores from younger raters.

Rater proficiency may also relate to how closely raters follow scoring criteria. Wolfe and colleagues (1998) defined proficient raters as those who 1) focus only on essay features explicitly described in the scoring criteria; 2) can understand and apply the scoring criteria in a general way, rather than focusing on individual essay features; 3) rely on the rubric to frame their scoring process; and 4) can handle high cognitive demand and consider an essay as a whole, rather than breaking the essay down into small pieces to evaluate. Wolfe and colleagues (1998) found that more proficient raters tend to follow the scoring criteria closer than less proficient raters. As such, it seems that it may be desirable to select more proficient raters to rate student work, as their ratings may better reflect the intended interpretation and use of the rubric compared to those of less proficient raters.

As previously mentioned, expert ratings may be collected from content experts in order to compare raters' scores to expert ratings that should theoretically align with the

intended interpretation and use of the rubric criteria. The extent to which non-content expert raters can provide adequate scores for performance assessments is mixed. Schoonen, Vergeer, and Eiting (1997) found that non-expert raters (i.e. raters with no specific training in the content area) provided less reliable scores than expert raters (i.e. raters with educational training and professional experience in the content area) when assessing writing ability. However, Powers and Kubota (1998) found that, after training, non-expert raters provided scores within an acceptable range of accuracy and could be interchangeable with expert raters. Consequently, educators and researchers may consider the type of training they offer to raters and think critically about whether the training is capable of adequately guiding raters who are unfamiliar with the content of the assessment.

In fact, rater training is often cited as an avenue by which to mitigate rater effects. Some researchers suggest that implementing and improving rater training mitigates rater effects (Borman, 1975; Elder, Knoch, Barkhuizen, & von Randow, 2005; McIntyre, Smith, & Hassett, 1984). However, others suggest that rater training may not mitigate rater effects to the desired levels (Engelhard, 1992; Lumley & McNamara, 1993; McNamara, 1996; Weigle, 1998). For example, implementing a rater training was found to successfully increase consistency of individual raters' scores, but was not found to adequately decrease differences in severity between new and old raters (Weigle, 1998). Though the relationship between rater training and rater effects is often discussed in the literature, detailed examples of rater training are lacking, and protocols for effective training remain largely unknown. Lack of training examples is problematic, especially considering that many researchers (e.g. Congdon & McQueen, 2000; Eckes, 2008;

Myford & Wolfe, 2004; Schaefer, 2008; Wu & Tan, 2016) suggest the presence of rater effects has implications for rater training.

It is important to remember it is unlikely that rater effects will ever be eliminated (Cronbach, 1990; McNamara, 1996; Wu & Tan, 2016). When rater effects are present in performance assessment scores, raters are not exchangeable with one another (Bejar, 2012). However, raters are often assumed to be interchangeable (Lunz et al., 1990). When raters are not interchangeable, the students' scores depend on which rater rated their products. A lack of exchangeability among raters is particularly problematic in the case of criterion-referenced assessments. The goal of a criterion-referenced assessment is to accurately place students into categories, typically regarding their proficiency with a particular construct (Crocker & Algina, 1986). However, if students' scores depend upon the rater, inaccurate classification decisions may be made regarding students' proficiencies (Wolfe, 2004; Wu & Tan, 2016). Thus, it is important that appropriate steps be taken to ensure exchangeability of raters. A first step is the evaluation of scores for rater effects.

**Evaluating Scores for Rater Effects**

When evaluating scores for rater effects, it is first important to remember that the absence of rater effects does not indicate score accuracy (Murphy & Balzer, 1989). Rater effects analyses simply allow researchers to evaluate the patterns present within ratings. The patterns of scores could be similar across raters and students, but the scores may not reflect students' actual abilities. For example, all raters may interpret the rubric similarly, resulting in similar score patterns across raters and students, but their interpretation of the rubric could be incorrect, resulting in inaccurate scores.

Moreover, the presence of rater effects does not necessarily indicate score inaccuracy (Wolfe, 2004). Because many rater effects analyses allow for the evaluation of score patterns in relation to a selected pool of raters, the prominence of rater effects is dependent upon the raters in the sample. Consequently, it could potentially be the case that "good" raters appear to have drastic rater effects if they are compared to a sample of "poor" raters, even though the "good" raters exhibit the most accurate rating tendencies (Wolfe, 2004). Thus, rater effects analyses should be interpreted cautiously and in the context of multiple sources of evidence.

Additionally, it could be the case that there is evidence of rater effects, but the scores exhibit a pattern accurate for students' abilities. For example, a rater may consistently assign low scores to student products, suggesting a severity effect. However, it could be the case that the rater happened to receive products from students of low ability, thus warranting a similar pattern of low scores across students. Or, a rater may assign similar scores across rubric elements, suggesting the presence of a halo effect. However, it could be the case that the student is of similar ability on all rubric elements, thus warranting similar scores across rubric elements. Or, a rater may assign moderate scores to student products, suggesting a restriction of range/central tendency effect. However, it could be the case that the students are of moderate ability, thus warranting a pattern of scores at the mid-point of the scoring levels across students.

Thus, in rater effects research, it is helpful to know the score most accurate for students, given their abilities and the scoring criteria. However, a limitation of much rater effects research is that the most accurate scores are often unknown (Engelhard, 1996; Wolfe, 2004). Researchers generally approach this issue in two ways: obtain expert rater

scores that are thought to represent the most accurate score for each student (e.g. Engelhard, 1996), or use statistical modeling techniques that allow researchers to glean an expected score for each student than can represent the most accurate score (e.g. Wolfe, 2004; Wu & Tan, 2016).

The Many-Facets Rasch Measurement (MFRM; Linacre, 1989) model allows researchers to glean students' expected scores and has been proposed for the evaluation of rater effects in performance assessment scores (e.g. Eckes, 2015; Engelhard, 1992, 1994; Myford & Wolfe 2003). The MFRM model allows for the inclusion of facets, or sources of variability thought to influence students' scores (Eckes, 2009). To obtain estimates of the extent to which rater effects are present in student scores, researchers can include a rater facet in the MFRM model. Inclusion of the rater facet allows for statistical tests and effect size measures that indicate variability in rater harshness/leniency or central tendency (Myford & Wolfe, 2004). Researchers can also include a rubric element facet to evaluate how the entire rubric or individual rubric elements function. Inclusion of the element facet allows for statistical tests and effect size measures that indicate variability across dimensions, suggesting whether or not a halo effect is present (Myford & Wolfe, 2004).

The MFRM model produces model-implied scores, which are estimated based on all facets present in the model and are thought to be invariant across raters (Engelhard, 1992). That is, the model-implied score is thought to represent the score a student should have received if rated by a rater of average leniency/severity. Students' model-implied scores are produced by taking into account how individual raters may have influenced the students' score (Stemler, 2004). Moreover, because the MFRM model provides estimates

of students' scores, researchers can compare students' model-implied scores and the raw scores students actually received from the rater. Researchers may then use the MFRM-generated model-implied score to statistically adjust raw scores generated by severe or lenient raters (Eckes, 2005; Wu & Tan, 2016). The MFRM model has gained popularity due to its versatility in the ability to include rater and rubric element facets. However, the MFRM model is not the only statistical technique that can be used to evaluate rater effects.

Generalizability theory (g-theory; Shavelson & Webb, 1991) may be used to evaluate scores for rater effects, and g-theory is perhaps one of the most common methods of evaluating the psychometric quality performance assessment scores. To appreciate g-theory, a brief interlude to classical test theory (CTT) and its shortcomings is necessary. In a traditional CTT framework, assessment scores are thought to be composed to two parts: "true" score variability and error variability (Haertel, 2006). In CTT, all error is considered to be unsystematic and all systematic error is considered to be "true" score variability. However, as mentioned, rater effects result from systematic variability in raters' ratings. Thus, it is possible that systematic errors due to raters are confounded with students' "true" score variability in a CTT framework. Thus, CTT has limitations when it comes to accurately determining the proportion of score variability due to differences in students' abilities. Moreover, because error is considered to be one lump sum of unsystematic variance in a CTT framework, CTT is not useful for identifying sources of systematic error variability. That is, researchers cannot parse out whether error results from differences in raters, differences in performance assessment tasks, or differences due to testing occasion (Bandalos, 2018).

G-theory provides a unique solution to the shortcomings of CTT. Specifically, in a g-theory framework, systematic error variability can be decomposed into different variance components via an analysis of variance (ANOVA; Shavelson & Webb, 1991). By computing variance components from assessment scores, systematic error variability due to facets such as rater, assessment task, rubric element, testing occasion, and their interactions can be parsed out. Variability related to each facet may then be compared to identify which facet(s) contribute the most systematic variability to scores. Researchers may specify relevant facets to be included in a g-theory analysis. For example, consider a performance assessment system in which students respond to a single performance task and all tasks are scored by the same two raters using a rubric with five dimensions. In this design, score variability may be decomposed into the proportion of total variability due to differences in student ability (e.g. student facet as the object of measurement) and two error facets: 1) proportion of total variability due to differences in raters (e.g. rater facet), and 2) proportion of total variability due to differences in rubric element difficulty (e.g. element facet). The decomposition of error variability into specific variance components can be useful to identify evidence of rater effects.

For example, if there is a main effect due to raters, it suggests that there may be differences in raters' leniency/severity. Though information from g-theory analyses can be helpful for identifying the presence of rater effects, results cannot indicate which individual raters are problematic. That is, g-theory provides evidence of rater effects at the group level, which is not particularly useful if researchers want to identify individual raters and provide feedback or recalibration for those raters. Because the courses of action to mitigate rater effects may vary depending on whether the effect occurs at the

group or individual level, it is important to evaluate scores for both group and individual rater effects.

For example, it is possible that, as a group, raters exhibit a central tendency effect. In this instance, the group-level effect could be due to issues with the rubric that result in raters being unable to differentiate between scoring criteria (Myford & Wolfe, 2004). As such, an appropriate avenue of action may be to reevaluate the rubric and make scoring criteria clearer, perhaps by reducing the number of scoring levels so raters can more easily differentiate between them. A group-level effect could also indicate the need for a more detailed rater training.

Conversely, evidence of individual-level rater effects may warrant further training with only a few raters. By identifying only certain raters to train further, researchers can save resources and train only specific raters, rather than the entire rater pool. Moreover, if there are individual-level effects, there may be implications for selection of raters. For example, if there is only evidence of rater effects with raters who are non-content experts, then researchers may use this information to justify selecting only content expert raters. Finally, evaluating the scores for individual rater effects is important because group-level analysis may "wash out" individual rater effects (Myford & Wolfe, 2003). Thus, there are many benefits to evaluating scores for individual rater effects. However, as discussed, g-theory is unable to detect individual rater effects. Instead, the MFRM model is proposed in order to obtain results for individual raters (Myford & Wolfe, 2003; Sudweeks, Reeve, & Bradshaw, 2005). In this study, an MFRM approach was used to evaluate Madison Collaborative ethical reasoning scores for group- and individual-level rater effects.

**Study Purpose & Research Questions**

This study served several purposes. First, no known studies have evaluated the presence of rater effects in the domain of ethical reasoning. Thus, this study serves as a contribution to the ethical reasoning literature. Second, given that students' ER-WR scores are used to make institution-level inferences regarding students' ethical reasoning abilities, it is important that scores are psychometrically sound and backed with evidence to support their interpretations and uses. Thus, a second purpose of this study was to determine the extent to which individual- and group-level rater effects influenced first- and second-year students' ER-WR scores. Results provide useful information for the Madison Collaborative in regard to rater selection and training. Finally, given that research suggests raters' content knowledge of the assessment may be related to rater effects, this study serves to evaluate whether there was a relationship between raters' knowledge of the 8KQs and their leniency or severity.

In this study, the following research questions were addressed:

1) Are there statistically significant differences in rater leniency/severity, suggesting a group-level leniency/severity rater effect?
2) Are there statistically significant rater leniency/severity and ER-WR rubric element interaction effects?
3) Is there a lack of distinguishability between score levels, suggesting a restriction of range effect? Is this restriction of range suggestive of a central tendency effect?
4) Is there a relationship between raters' knowledge of the 8KQs and rater leniency/severity?

# Chapter 3: Method

## Participants

**Student participants.** Student participants were first- and second-year students at James Madison University. Essays were collected from 484 students, with 330 essays from first-year students and 154 essays from second-year students.

**Raters.** Eighteen raters were recruited to rate ER-WR essays. All raters were employed at James Madison University at the time of rating. Raters were recruited from the academic affairs and student affairs divisions, with representation from several colleges and student affairs offices. All raters were familiar with the 8KQs prior to rating; however, experience with the 8KQs varied.

## Measures

**Ethical Reasoning and Writing (ER-WR) essay assessment.** The ER-WR essay assessment consists of the ER-WR essay prompt and the ER-WR rubric. The ER-WR essay assessment is a performance assessment on which students are asked to describe 1) an ethical situation with which they were familiar, 2) the ethical considerations relevant to the situation, 3) their ethical reasoning process, and 4) the decision they made (See Appendix A for ER-WR instructions and prompt). The ER-WR essay prompt was developed by an ethical reasoning expert on campus. The ER-WR rubric (see Appendix B) was used by trained raters to score students' ER-WR essays. The ER-WR rubric was developed jointly by an assessment expert and ethical reasoning expert on campus. Scores range from 0 – 4 on five elements designed to encompass the five elements of students' ethical reasoning processes when using the 8KQ framework. As previously discussed, the five elements are thought to be sequential in the ethical reasoning process.

***Element A.*** Element A is labeled "Ethical situation: Identifying an ethical issue in its context." In this element, students are rated on their abilities to 1) identify and describe an ethical situation they have faced, and 2) describe relevant contextual features surrounding the situation. Element A is the first element on the rubric because, before students can evaluate an ethical dilemma, they must be able to delineate between an ethical decision and a difficult decision. Most ethical decisions are difficult, but not all difficult decisions are ethical in nature. To receive credit for Element A, students must delineate between difficult and ethical situations. Essays in which students do not provide an ethical situation are considered unrateable and do not receive scores for any ER-WR rubric elements. If a student provides an ethical situation, the distinguishing factors between scores on Element A are related to the ability of the student to 1) explicitly describe the potential decision options and 2) describe the relevant details of the ethical situation.

***Element B.*** Element B is labeled "Key question reference: Mentioning the 8KQs or equivalent terms." After students identify an ethical situation, they must identify relevant considerations. Scores on Element B are directly related to how many KQs students explicitly reference in their essays. Explicit references include the direct mention of the KQs by name (e.g. Fairness, Outcomes, Responsibility, Character, Liberty, Empathy, Authority, and Rights). Students can also implicitly reference the KQs in their essays by using synonyms or phrases that get at the gist of the KQs. However, if students only implicitly reference KQs, they cannot receive above a score of one on Element B.

***Element C.*** Element C is labeled "Key question applicability: Describing which of the 8KQs are applicable or not applicable to the situation and why." After students

identify relevant considerations, they must provide a rationale for the applicability or (in)applicability of considerations to their ethical situations. Scores on Element C are directly related to the number of KQs for which students provide a rationale for their (in)applicability to their ethical situation. Theoretically, Element C builds upon Element B, as students must mention the KQs to be able to provide a rational for their (in)applicability to their ethical situation. Thus, students cannot score higher on Element C than Element B. However, students may score lower on Element C than Element B, as they can merely mention KQs to receive credit for Element B, but students must provide a rationale for the (in)applicability of each of those KQs to their ethical reasoning situation to receive credit for Element C.

***Element D.*** Element D is labeled "Ethical reasoning: Analyzing individual KQs." After students identify which KQs are relevant to their ethical situations, they must analyze the KQs within the context of their ethical situations. Thus, in this element, students are rated on their abilities to analyze the KQs in the context of their ethical situations. Element D is theoretically tied to Element C in the sense that students must reason through which KQs are applicable before they can effectively analyze the KQs in the context of their ethical situations. Element D is also empirically tied to Element C through a "special note" on the rubric that indicates students must identify three or more applicable KQs (i.e. receive a score of 1.5 or higher on element C) to achieve a score higher than one on element D.

***Element E.*** Element E is labeled "Ethical reasoning: Weighing the relevant factors and deciding." After students analyze the KQs in the context of their ethical situations, they must come to a decision regarding their ethical situation. Thus, in this

element, students are rated on their abilities to weigh the KQs and other relevant factors to come to decisions regarding their ethical situations. Element E is theoretically tied to Element C in the sense that students cannot provide a rationale for the applicability of at least three KQs, then they cannot balance KQs to come to a logical ethical decision. Thus, similar to the relationship between Elements C and D, Element E is also empirically tied to Element C through a "special note" on the rubric that indicates students must identify three or more applicable KQs (i.e. receive a score of 1.5 or higher on Element C) to achieve a score higher than one on Element E. Moreover, Element E is theoretically tied to Element D, as if students are not able to accurately analyze at least three KQs, then they cannot effectively weigh the KQs to come to a logical decision. Thus, Element E is empirically tied to Element D through a "special note" on the rubric that indicates students must accurately analyze three or more key questions (i.e. receive a score of 1.5 or higher on Element D) to achieve a score higher than one on Element E.

Typically, students score highest, on average, on Element A and second-highest, on average, on Element B. Elements C, D, and E typically yield the lowest scores, on average, across the five rubric elements. On average, across all rubric elements, first-year students typically score higher than second-year students, likely because first-year students take the ER-WR the day after participating in *It's Complicated*. Since 2013, first-year students' average scores across all rubric elements have ranged from 1.11 to 1.51, and second-year students' average scores across all rubric elements have ranged from 0.88 to 1.21.

**Ethical Reasoning Identification Test (ERIT).** The ERIT is a 50-item multiple choice assessment typically administered to students to measure their abilities to identify

relevant KQs when provided with a brief scenario. Specifically, the ERIT consists of 42 multiple choice items on which examinees are asked to select the KQ most relevant to ethical scenarios, as well as two testlets, with each testlet having four items related to one ethical scenario. On all items, students are provided eight response options (Fairness, Outcomes, Responsibilities, Character, Liberty, Empathy, Authority, and Rights) from which to choose.

Confirmatory factor analyses suggested that a unidimensional model provided adequate fit to student scores (Bashkov et al., 2014; Holzman et al., 2017; Smith et al., 2015; Smith et al., 2016). Cronbach's alpha as a measure of internal reliability has been above 0.79 each year (Bashkov et al., 2014; Holzman et al., 2017; Smith et al., 2015; Smith et al., 2016), suggesting adequate internal consistency reliability for student scores. The ERIT has never been administered to populations other than undergraduate students. Thus, psychometric information for scores from other populations, such as the faculty/staff raters in the current study, is unavailable.

In this study, only the first 42 items were administered to raters. Scores were a proxy of raters' knowledge of the 8KQs. Possible scores ranged from 0 – 42, with higher scores representing more knowledge of the 8KQs than lower scores. The average score on the ERIT was 36.17, with a standard deviation of 3.54. As only eighteen raters participated in this study, sample size was not large enough to conduct a confirmatory factor analysis to evaluate the factor structure of scores. As a form of validity evidence, Cronbach's alpha was estimated. Cronbach's alpha was 0.65 for this sample of raters.

## Procedure

**ER-WR essay collection.** All essays were collected on university-wide assessment days. Assessment day is a proctored, low-stakes, standardized testing occasion during which students are administered a battery of cognitive and non-cognitive assessments that can be completed within two-hours. Assessment day is designed for longitudinal data collection. Thus, students will take the same assessments as second-year students that they took as first-year students. However, note that all data in this study were cross-sectional. That is, data were from the 2017-2018 academic year, with a cohort of first-year students assessed in August of 2017 and a different cohort of second-year students assessed in February 2018.

First-year students completed the ER-WR assessment in August 2017 prior to beginning first-semester courses. On the day before taking the assessment, first-year students experienced *It's Complicated*. Second-year students completed the ER-WR assessment in February 2018 after completing 45 – 70 credit hours. Second-year students experienced *It's Complicated* as first-year students and may or may not have had Madison Collaborative or 8KQ interventions in their coursework and/or co-curricular activities.

Using campus computer labs, all essays were written electronically on a university-developed testing platform. Students were granted 55 minutes to complete the ER-WR essay. Trained proctors walked throughout the testing rooms, encouraging students to take the full amount of time, check their work, and expand upon their essays. Though there is no minimum word count for the ER-WR essay, students were encouraged to write no fewer than 250 words.

**Rating.** After rating, all raters completed an IRB-approved informed consent to allow their scores to be analyzed and reported in various contexts, both within and external to James Madison University.

The rating process occurred over two days in May 2018 (see Appendix D for rating timeline). Breakfast and lunch were provided each day. First-time raters were remunerated at a rate of $250/day and returning raters were remunerated at a rate of $300/day. Raters were placed into anonymous rater teams, ensuring that all essays were rated by two raters. Historically, raters indicated that they feel fatigued 1) by the end of each rating day, and 2) by the end of the entire rating session. Thus, to mitigate rater effects due to fatigue, essays were counterbalanced within rater pairs. That is, raters within each rater team rated essays in reverse order (e.g. the first essay rater one rated was the last essay rater two rated).

All student essays were de-identified prior to rating. Historically, the Madison Collaborative has rated both first- and second-year student essays in the same rating session. The Madison Collaborative justifies the combined rating on two accounts: 1) raters are unaware as to which essays were written by which students, and thus do not know which essays are from first-year students and which essays are from second-year students; and 2) though first- and second-year student essays may vary in quality, the ER-WR rubric was designed with a wide range of scoring criteria, so the same rubric can be used across essays of varying ethical reasoning quality and with college students of different ages.

***Day one.*** All raters participated in a two-hour rater training. The rater training was conducted by one facilitator, who was a quantitative psychology faculty member and

assessment liaison to the Madison Collaborative. This facilitator was selected to maintain consistency with previous MC rater trainings, as she was the primary rater training facilitator in previous years. During rater training, raters were introduced to the ER-WR rubric and the 8KQ synonyms (see Appendix E) deemed acceptable by ethical reasoning experts on campus. After raters were introduced to the ER-WR rubric and 8KQ synonyms, raters rated two practice essays.

Prior to training, each practice essay was rated by an ethical reasoning content expert using the ER-WR rubric. The expert provided scores and a rationale for the scores for each ER-WR rubric element. In the training, the facilitator used the expert rater's scores and rationales to guide raters as they rated the practice essays. The first practice essay was of excellent ethical reasoning quality (e.g. received a score of about three or higher on each ER-WR rubric element). This particular practice essay was chosen to demonstrate to raters the qualities of a high-scoring essay. A high scoring essay was selected because, historically, raters participating in MC rating sessions indicated not having an adequate conception of high-quality ethical reasoning essays from the training process. Raters indicated that, when they rated students' essays, they tended to provide scores that were higher than appropriate because they did not have an example of what high-quality ethical reasoning looks like in the training. To improve raters' conceptions of the skills that warrant high scores on the ER-WR rubric, a high scoring essay was selected.

Individually, raters scored Element A. Guided by the facilitator, raters then had large-group discussion regarding the Element A score. Several raters shared their scores and rationale for their scores. The facilitator then shared the expert rating for Element A

and discussed why the expert rating was the most appropriate score. Where necessary, the facilitator assisted raters in making distinctions between scoring criteria. This same process ensued for ER-WR rubric Elements B – E. For all elements, raters were encouraged to calibrate within a half-point of the expert rater (i.e. if the expert rater provided a score of 2, raters were instructed that scores between 1.5 and 2.5 were acceptable). This half-point calibration and calibration process resembled trainings from prior years. Note that raters did not submit scores for the facilitator to check that they were calibrating to the expert rater. It was assumed that raters recognized the logic behind the expert rater's scores and would attempt to mimic the same logic when scoring students' essays.

The second practice essay was of good ethical reasoning quality (i.e. an average score of about two on the ER-WR rubric). This practice essay was of lower ethical reasoning quality than the first practice essay and was selected to provide raters practice with distinguishing between the middle score levels of the ER-WR rubric. Raters were allotted approximately fifteen minutes to rate all ER-WR rubric elements. Raters then discussed their ratings in pairs or small groups, based on where they were sitting. To ensure that raters did not discuss their ratings with their anonymous rater pair, raters were assigned seats. Seat assignment was important because if raters compared with their partner, they ran the risk of calibrating to their partner, rather than the expert rater. If rater pairs calibrate to one another, students' scores may become biased (e.g. if both raters become harsh, the student will receive a lower score than is warranted, given the student's ability, and scores will not balance across raters). Ratings were then discussed as a large group. The facilitator shared the expert rating for each rubric element and

discussed why the expert rating was most appropriate. Where necessary, the facilitator assisted raters in making distinctions between scoring criteria. Just as with the first training essay, raters were encouraged to calibrate within a half-point of the expert rater.

After the training, raters completed the ERIT. Recall, the ERIT was administered as test of raters' 8KQ knowledge. The ERIT was completed via a paper-and-pencil scantron form. All raters completed the ERIT prior to beginning the rating process. After completing the ERIT, raters were given lunch. After lunch, three hours remained, during which raters began rating student essays. Over the course of days one and two, sixteen raters were assigned 59 essays to rate and two raters were assigned 57 essays to rate. Raters' assigned essays included first- and second-year student essays. As previously mentioned, first- and second-year student essays were randomly dispersed among the raters' essays and raters were unaware as to which essays were written by first-year students and which essays were written by second-year students.

As part of raters' assigned essays, five plant essays were administered to each rater. Thus, raters rated either 52 or 54 essays unique to their rater pair, and five essays that were common across all raters. The plants were administered in the same order for all raters (i.e. plant 1 was administered as the $6^{th}$ essay raters would rate, plant 2 was administered as the $12^{th}$ essay raters would rate, plant 3 was administered as the $18^{th}$ essay raters would rate, plant 4 was administered as the $24^{th}$ essay raters would rate, and plant 5 was administered as the $30^{th}$ essay raters would rate). Plants were administered to create links across raters, which was necessary for the analyses described later in this chapter. Without common essays rated by all raters, model parameters cannot be

calibrated together, resulting in an inability to compare the parameters necessary to evaluate the research questions of this study (Eckes, 2009, 2015; Linacre, 2017a).

In the three hours of rating on day one, raters completed as many essays as they were able, without an expectation of the number of essays they would rate. Raters returned on day two to complete their remaining essays.

***Day two.*** On day two, raters participated in a one-hour refresher training for the ER-WR rubric. Individually, raters were allotted approximately fifteen minutes to rate one practice essay of excellent ethical reasoning quality. Similar to the rationale for selecting an essay of high ethical reasoning quality on day one, a high quality essay was selected to remind raters of the essay characteristics that warrant the highest scores on the ER-WR rubric. Raters then discussed their ratings in pairs or small groups. Like day one, seating was assigned to ensure that raters did not compare practice ratings with their partner. Ratings were discussed as a large group, with the facilitator sharing the expert rating for each rubric element. The facilitator also discussed why the expert rating was most accurate. Where necessary, the facilitator assisted raters in making distinctions between scoring criteria. Just as in day one, raters were encouraged to calibrate within a half-point of the expert rater.

After rating the practice essay, raters continued to rate in the same manner as they did on day one. As such, raters remained in the same rater pairs as day one, and they began rating the essay on which they left off the previous day. Raters had two hours to rate essays before lunch. Raters were then allowed a thirty-minute lunch break. After lunch, raters had four hours to complete their assigned essays. All raters completed their assigned essays by 2pm on the second rating day.

## Data Analysis

Unless otherwise stated, all data screening and preparation was conducted using SAS Software Version 9.4 (SAS Institute, 2015). Unless otherwise stated, all data analysis was conducted using FACETS (Linacre, 2017b).

**Data screening.** First, data were screened to remove essays that were considered "unrateable" because the student did not present an ethical dilemma. These essays were identified by raters commenting that the essays were not ethical dilemmas. I reviewed the essays that were flagged for not having an ethical dilemma, and if an ethical dilemma was not present, scores for that essay were removed from the analysis. If an ethical dilemma was present, the scores were retained. Twenty essays did not have an ethical dilemma present, resulting in 464 remaining student essays, and a total of 4,640 scores in addition to the 450 scores from the plant essays.

Next, data were screened to ensure that raters used the special notes on the ER-WR rubric. Because ER-WR rubric developers consider use of the special notes necessary for generating valid scores, any scores in which the raters did not use the special notes (i.e. assigned a score of 1.5 or higher on elements D and/or E when assigning a score of 1 or lower on element C), were removed from the analysis. There were eight instances in which raters did not follow the special notes, resulting in a total of 4,560 ratings in addition to the 450 ratings from the plant essays. In total, there were 5,010 ratings for analysis.

Next, data were screened for missingness. When screening for missingness, I checked to ensure that all raters completed at least one of the five plant essays. Because the plants are necessary to facilitate comparisons of parameters across raters (Eckes,

2009, 2015; Linacre, 2017a), the data were screened specifically to identify any raters who had not completed at least one of the five plant essays. All raters completed all five plant essays. No other missing data were present.

**Data preparation.** First, a total score was computed for the ERIT in order to evaluate research question 4. To compute the total score, first raters' responses to the ERIT questions were scored correct/incorrect, where incorrect scores were provided a score of zero and correct scores were provided a score of one. Next, raters' scores were summed across all ERIT items, yielding a single total score thought to be a proxy for raters' knowledge of the 8KQs.

FACETS (Linacre, 2017b) requires data to be in integer form. Because raters could provide half-point scores for all ER-WR rubric elements, all ER-WR scores were multiplied by two. Thus, analyzed scores ranged from 0 – 8.

**Many-Facets Rasch Measurement.** All research questions were evaluated using Many-Facets Rasch Measurement (MFRM; Linacre, 1989). The MFRM model features two key advantages to researchers. First, all facets are placed on the same logit measurement scale, allowing for comparisons to be made across facets (Bond & Fox, 2015). Second, the MFRM model provides model expected estimates of the scores students should have received, after accounting for measurement error related to all included facets. Researchers may compare model-implied scores and raw scores to make inferences regarding the extent to which rater-assigned raw scores represent students' scores after correcting for measurement error (Eckes, 2009; Engelhard, 1994; Wu & Tan, 2016).

The MFRM model is an extension of the single-facet rating scale model (Andrich, 1978) and single-facet partial-credit model (Masters, 1982) and allows for multiple facets to be included in the evaluation of polytomously-scored assessment items. Specifically, rater and rubric element facets can be included to evaluate performance assessment scores. With student, rater, and rubric element facets, a rating scale model may be defined as

$$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_k, \quad (1)$$

where $P_{nijk}$ is the probability of student *n* being rated *k* on element *i* by rater *j*, $P_{nijk-1}$ is the probability of student *n* being rated *k*-1 on element *i* by rater *j*, $\theta_n$ is the ability of student *n*, $\delta_i$ is the difficulty of ER-WR rubric element *i*, $\alpha_j$ is the severity of rater *j*, and $\tau_k$ is the difficulty of score level *k* compared to score level *k*-1 (Eckes, 2015; for a list of equations, see Appendix F). When a rating scale model is specified, the researcher assumes that all raters use the set of rubric elements in the same way when rating. When a rating scale model is specified, all rubric elements must also have the same number of score levels (Bond & Fox, 2015; Myford & Wolfe, 2003).

With student, rater, and rubric element facets, a partial credit model may be defined as

$$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_{ijk}, \quad (2)$$

where $P_{nijk}$ is the probability of student *n* being rated *k* on element *i* by rater *j;* $P_{nijk-1}$ is the probability of student *n* being rated *k*-1 on element *i* by rater *j;* $\theta_n$ is the ability of student *n*, $\delta_i$ is the difficulty of ER-WR rubric element *i;* $\alpha_j$ is the severity of rater *j;* and $\tau_{ijk}$ is the difficulty of score level *k* compared to score level *k*-1, which is free to vary

across ER-WR rubric element *i* and rater *j* (Eckes, 2015).When a partial credit model is specified, the researcher assumes each rater uses each rubric element in their own individual ways. Thus, the partial credit model is a more complex model than the rating scale model and allows for the estimation of additional parameters for both raters and rubric element thresholds (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003).

Regardless of whether a rating scale or partial credit MFRM model is used, the log-odds of students obtaining scores of *k* are a function of the additive effects of their abilities, the difficulty of the ER-WR rubric element, rater severity, and the difficulty of scoring in score level *k* compared to *k*-1 (Eckes, 2009, 2015; Linacre, 2017a; Myford & Wolfe, 2003). In this study, variations of a hybrid of equation 1 and equation 2 were used. The hybrid MFRM model for this study may be defined as

$$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_{ik}, \quad (3)$$

where $P_{nijk}$ is the probability of student *n* being rated *k* on element *i* by rater *j*, $P_{nijk-1}$ is the probability of student *n* being rated *k*-1 on element *i* by rater *j*, $\theta_n$ is the ability of student *n*, $\delta_i$ is the difficulty of ER-WR rubric element *i*, $\alpha_j$ is the severity of rater *j*, and $\tau_{ik}$ is the difficulty of score level *k* compared to score level *k*-1 for ER-WR rubric element *i* (Eckes, 2015).

All MFRM models were estimated using joint-maximum likelihood estimation via FACETS 3.80.0 (Linacre, 2017b). The model used for each research question is defined below with each respective research question. Upon analyzing the data using the appropriate model, various indices were used to evaluate each research question. These indices are commonly used in the literature to evaluate rater-mediated scores for rater effects (e.g. Engelhard, 1992, 1994; Eckes, 2005, 2008; Weigle, 1998; Wu & Tan, 2016).

First, a brief overview of each index and its computation is provided. Note that each index is provided in FACETS (Linacre, 2017b) output. However, equations are provided for the benefit of the reader. Specific interpretations and ideal outcomes for each index are provided with each research question, where appropriate.

**Fixed-effect chi-square.** The fixed-effect chi-square is a significance test used to test the null hypothesis that there are no differences in the logit values for an object of measurement (e.g. student, rater, ER-WR element), after controlling for measurement error (Eckes, 2015; Myford & Wolfe, 2003). For example, a non-significant chi-square for students suggests that all students exhibit the same ability, after controlling for measurement error. In this study, the object of measurement was either student or rater. The fixed-effect chi-square is defined as

$$x^2 = \sum(w_o * D_o^2) - \frac{(\sum w_o * D_o)^2}{\sum w_o}, \quad (4)$$

where $D_o$ is the estimated logit of the object of measurement (i.e. difficulty of ER-WR rubric element, severity/leniency of rater, or student ability) and $w_o = \frac{1}{SE_o^2}$ (Myford & Wolfe, 2003). Degrees of freedom equal $L - 1$, where $L$ = the number of observations of the object of measurement (Myford & Wolfe, 2003). Note that the fixed-effect chi square is sensitive to sample size. Thus, in large samples, the fixed-effect chi square may be statistically significant, even with small differences in the object of measurements' logits (Eckes, 2015).

**Separation ratio.** Note that the separation ratio will not be reported directly to evaluate research questions; however, it is described because it provides the foundation upon which subsequent indices are computed. The separation ratio quantifies the precision of the spread of the logits associated with the object of measurement in relation

to the measurement error associated with the object of measurement's logit values (Eckes, 2015; Myford & Wolfe, 2003). Said differently, the separation ratio indicates how precisely the object of measurement is able to be spread across the logit continuum. The separation ratio requires the computation of the *true SD*, defined as

$$SD_t^2 = SD_o^2 - MSE, \tag{5}$$

where $SD_o^2$ is the standard deviation of the observed logits for a given object of measurement, and $MSE$ is the average measurement error associated with a given object of measurement (Eckes, 2015). The separation ratio ($G_o$) may then be defined as

$$G_o = \sqrt{\frac{SD_t^2}{MSE}}. \tag{6}$$

$G_o$ ranges from 0 to positive infinity, with values near 0 indicating less spread of the object of measurement across the logit continuum, compared to higher values (Eckes, 2015; Myford & Wolfe, 2003).

**Separation index.** The separation index ($H_o$) is an extension of the separation ratio and is defined as

$$H_o = \frac{4\sqrt{\frac{SD_t^2}{MSE}}+1}{3}. \tag{7}$$

$H_o$ ranges from 0 to positive infinity and indicates the number of statistically significantly different levels there are of the object of measurement (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003). For example, an $H_o$ of 5.3 suggests that five distinct strata of the object of measurement exist. A value of $H_o$ near 1.0 suggests that only one strata of the object of measurement is distinguished (Eckes, 2015).

**Reliability of separation.** The reliability of separation ($R_o$) is also an extension of the separation ratio and is defined as

$$R_o = \frac{\frac{SD_t^2}{MSE}}{1+\frac{SD_t^2}{MSE}}. \quad (8)$$

$R_o$ ranges from 0.0 to 1.0 and is analogous to traditional reliability indices, such as Cronbach's alpha (Myford & Wolfe, 2003). Conceptually, similar to how Cronbach's alpha is an estimate of how reliably students can be separated along the ability continuum, the $R_o$ is an estimate of how reliably the object of measurement can be separated along the logit continuum. Higher reliability of separation values indicate more reliable separation of the object of measurement than lower values (Bond & Fox, 2015; Eckes, 2015). Moreover, the reliability of separation may be interpreted as the proportion of an object of measurement's observed score variability that is not due to measurement error (Eckes, 2015).

**Evaluation of MFRM assumptions.** Prior to analysis, three MFRM assumptions were evaluated: local independence, unidimensionality, and correct model form.

***Local independence.*** Local independence refers to the assumption that item responses are independent from one another after controlling for the construct of interest (DeMars, 2010). When local independence is violated, it may be that a secondary construct is measured with the item, or it may be that an item influences responses to subsequent items (Marais & Andrich, 2008). In this study, local independence would be met if students' probabilities of receiving a certain score on an ER-WR rubric element were not related to the score they received on a previous element, after controlling for students' ethical reasoning abilities. However, due to the special notes on the rubric, it was plausible that scores would not suffice the local independence assumption. Specifically, scores on Elements D and E were expected to be dependent on Element C, and scores on Element E were expected to be dependent on Element D.

Violations of local independence are problematic because they may influence parameter estimates (Li, Li, & Wang, 2010; Smith, 2005) as well as inflate reliability estimates (Marais & Andrich, 2008; Wainer & Thissen, 1996; Wang & Wilson, 2005). When local independence is violated, a common response is to sum the dependent items to create a single polytomous item (DeMars, 2010; Marais & Andrich, 2008; Stone & Zhu, 2015). Given that the purpose of this study was to evaluate the presence of rater effects in scores across all ER-WR rubric elements, it was not advantageous to sum students' scores on Elements C, D, and E to create a single polytomous rubric element. Thus, to address the likely violation of local independence, Elements D and E were each split into two elements, resulting in Elements D lower, D upper, E lower, and E upper (M. Linacre, personal communication, February 26, 2018). Scores were then assigned to respective upper or lower elements based on the special notes. The range of the upper and lower elements matched the range of scores available to students, based on the special notes. That is, Element D lower and E lower ranged from 0 – 1 (0 – 2 when transformed to integers for FACETS), as the special notes do not allow students to obtain a score higher than one on Elements D or E if they receive a score of one or lower on Elements C or D. Element D upper and E upper ranged from 0 – 4 (0 – 8 when transformed to integers for FACETS), as students are able to obtain scores across the full spectrum of score levels if they receive a score above one on Elements C or D.

As an example of how data were structured in FACETS (Linacre, 2017b), consider the three possible cases of scoring: 1) students receive above a score of one on Elements C and D, thus voiding the special notes; 2) students receive above a score of one on Element C, but receive a score of one or less on Element D, thus voiding the

special note for Element C but maintaining the special note for Element D; or 3) students receive a score of one or less on Element C, thus maintaining the special notes for Elements C and D and requiring scores of one or less on Elements D and E. In the first case, students received scores for Elements D upper and E upper, and data were considered missing for Elements D lower and E lower. In the second case, students received a score for Element D upper and E lower, and data were considered missing for Elements D lower and E upper. In the third case, students received scores for Elements D lower and E lower, and data were considered missing for Elements D upper and E upper. See figure 1 for an example of how the data were structured in FACETS for each case. Note that figure 1 is for illustrative purposes only and the rater facet was left out of figure 1 for simplicity.

To determine whether the local independence assumption was violated for these data, two models were run: 1) one analysis in which the MFRM hybrid model (equation 3) was specified for the data where the five-element structure (i.e. Elements A, B, C, D, and E) was maintained, and 2) one analysis in which the MFRM hybrid model (equation 3) was specified for the data where Elements D and E were each split into two elements (i.e. Elements A, B, C, D lower, D upper, E lower, and E upper). The first model will be referred to as the five-element structure model, and the second model will be referred to as the seven-element structure model. Given that local independence violations inflate reliability estimates, local independence was considered violated if the student reliability of separation index was at least 0.05 higher for the five-element structure than the seven-element structure (Marais & Andrich, 2008). Note that there are no thresholds presented in the literature for this comparison. However, Marais and Andrich (2008) found an

increase of at least 0.05 in the student reliability of separation when they introduced dependencies, so this value was used in this study as a threshold for determining whether local independence was violated.

Given that missingness is induced due to students' scores on Element C and/or Element D, a brief discussion of missing data is warranted. Specifically, because missingness was created due to students' scores on Element C and/or Element D, and students' scores for those elements were included in the analysis, missing scores for Element D lower, Element D upper, Element E lower, and Element E upper were considered missing at random (MAR; Enders, 2010). Joint-maximum likelihood estimation may be used with MAR data (Linacre, 2017b), mitigating concerns about biased results due to missingness.

***Unidimensionality.*** Unidimensionality is related to local independence and refers to the assumption that all assessment items measure only one, common construct (Bandalos, 2018; DeMars, 2010). Unidimensionality was evaluated by conducting a Principal Components Analysis (PCA) on the standardized residuals. The PCA was conducted using SAS Software Version 9.4 (SAS Institute, 2015). Standardized residuals were estimated via

$$Z_{nij} = \frac{x_{nij} - e_{nij}}{\sqrt{w_{nij}}} \tag{9}$$

where $x_{nij}$ is the observed rating for student *n* on element *i* assigned by rater *j;* $e_{nij}$ is the expected rating for student *n* on element *i* assigned by rater *j*, given the model; and $w_{nij}$ is the variability of the observed rating around its expected rating, given the model, otherwise known as model variance (Eckes, 2015).

The expected rating may be further defined as

$$e_{nij} = \sum_{k=0}^{m} k p_{nijk} \quad (10)$$

where *k* is a rating and $p_{nijk}$ is the probability of student *n* obtaining score *k* on element *i* from rater *j*, given a specified MFRM model (Eckes, 2015). The model variance may be further defined as

$$w_{nij} = \sum_{k=0}^{m} (k - e_{nij})^2 p_{nijk} \quad (11)$$

where all components are as defined in equation 10 (Eckes, 2015). The square root of model variance is the statistical information contributed by a particular rating (Myford & Wolfe, 2003).

PCA analyses in the Rasch framework are used to evaluate whether there are systematic patterns in the residuals. If there are patterns in the residuals, a secondary dimension, often referred to as a "contrast," may be present. It is assumed that all elements are grouped on the first contrast, and the PCA specifically tests whether any elements group on secondary contrasts ("Dimensionality: Contrasts and Variances," n.d.). Each contrast has an associated eigenvalue, and the eigenvalues represent the number of elements that make up the respective contrast. If eigenvalues for secondary contrasts are less than 2.0, indicating there are fewer than two elements on the secondary contrasts, then the researcher has evidence of unidimensionality ("Dimensionality: Contrasts and Variances," n.d.). In this study, unidimensionality was considered to be met if the eigenvalues for the secondary contrasts were less than 2.0.

***Correct model form.*** Correct model form refers to the idea that an appropriate model is used to analyze the data. Data will never fit the model perfectly (Linacre, 2003). However, fit indices may be used to determine whether the data fit the model enough to

yield estimates useful for evaluating research questions. Correct model form was evaluated in two ways: 1) overall model fit, and 2) rater fit.

*Overall model fit.* To evaluate overall model fit, the absolute value of the standardized residuals were evaluated. Standardized residuals indicate how many standard deviations the observed score deviated from the expected score. Given that standardized residuals of |2.0| indicate that the observed score deviated by two standard deviations from the expected score, standardized residuals greater than |2.0| indicate highly unexpected scores, as they would be expected to appear less than 5% of the time in data that are consistent with the chosen MFRM model (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003; Wright & Masters, 1982). Thus, data were thought to fit the model well overall if fewer than 5% of the standardized residuals were greater than or equal to |2.0|.

*Rater fit.* Because the primary object of analysis in this study is raters, rater fit was evaluated. To evaluate rater fit, the unweighted mean square ($MS_U$) and weighted mean square ($MS_W$) indices were evaluated. $MS_U$ is an average of raters' squared standardized residuals (equation 9) for all students and elements and is defined as

$$MS_{U_j} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} z_{nij}^2}{NI}, \tag{12}$$

where $N$ = the number of students rated by that rater, and $I$ = the number of elements (Eckes, 2015).

$MS_W$ is defined as

$$MS_{W_j} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} z_{nij}^2 w_{nij}}{\sum_{n=1}^{N}\sum_{i=1}^{I} w_{nij}}, \tag{13}$$

where all terms are as defined in equation 9 (Eckes, 2015). $MS_W$ values are weighted by statistical information, resulting in differential weighting of ratings. Specifically, ratings

assigned in score levels further from the examinees' ability are weighted less heavily than ratings assigned to the other score levels, as less information is contributed to the model by these extreme scores (Bond & Fox, 2015; Eckes, 2015). Thus, though $MS_U$ and $MS_W$ are similar, they provide slightly different information to researchers and both were evaluated in this study. Note that $MS_U$ and $MS_W$ may be referred to as Mean Square outfit and Mean Square infit, respectively, in other references (e.g. Bond & Fox, 2015; Eckes, 2005; Engelhard, 1994, 2002; Myford & Wolfe, 2003).

$MS_U$ and $MS_W$ range from 0 to positive infinity, with values of 1.0 indicating perfect fit of the data to the model (Linacre, 2003). Values less than 1.0 indicate that the observed ratings are more similar to the model-implied ratings than would be predicted by the model (i.e. overfit of the model), and values greater than 1.0 indicate that the observed ratings are less similar to the model-implied ratings than would be predicted by the model (i.e. underfit of the model; Eckes, 2015; Linacre, 2003). Note that $MS_U$ and $MS_W$ indices may be transformed to a *t*-distribution to test the statistical significance of perfect model-data fit (Eckes, 2015). Or, $MS_U$ and $MS_W$ may be left untransformed and used as effect sizes. Using $MS_U$ and $MS_W$ as indicators of both statistical significance and effect size is not common in Rasch measurement (DeMars, 2010). For purposes of this study, $MS_U$ and $MS_W$ were maintained as untransformed measures of effect size.

Researchers have proposed various benchmarks for acceptable fit. Linacre (2003) proposed that $MS_U$ and $MS_W$ measures between 0.5 – 1.5 are often accepted as indicators of acceptable fit. However, Bond and Fox (2015) suggested that narrower limits between 0.7 – 1.3 are appropriate. Given that use of scores is relatively low stakes for this study, $MS_U$ and $MS_W$ values between 0.5 and 1.5 were considered acceptable. Though there are

no hard benchmarks for acceptable $MS_U$ and $MS_W$ values, values above 2.0 are considered major distortions in model fit (Eckes, 2015; Linacre, 2003). As such, $MS_U$ and $MS_W$ values greater than 2.0 were flagged as indications of major rater misfit.

After assumptions were evaluated for each MFRM model, data were analyzed in accordance with each research question. In all analyses, facets were oriented such that greater logits for student ability represented more ability than lower logits, greater rater logits represented more severity in rating than lower logits, and greater element logits represented more difficulty than lower logits. The average logits of the rater and element facets were fixed to 0.00, and the average student ability logit was freely estimated. In the following section, I describe data analysis procedures and indices relevant to evaluate each research question.

**Research Questions**

**Research question 1: Are there statistically significant differences in rater leniency/severity, suggesting a group-level leniency/severity rater effect?** The hybrid MFRM model (equation 3) was used to evaluate this research question. First, the fixed-effect chi-square was evaluated as a global test of whether leniency/severity differed across raters. The fixed-effect chi-square (equation 4) was estimated to evaluate the null hypothesis that, after controlling for measurement error, there were no differences in rater severity. A statistically significant chi-square ($p<.05$) suggests that at least two raters are statistically significantly different in their leniency/severity logit scores (Myford & Wolfe, 2004).

Next, the rater separation index and reliability of rater separation were evaluated with raters as the object of measurement. In the rater separation ratio (equation 6), the

true SD was computed using the observed standard deviation of the rater logits and the standard error associated with the rater logits. Further, the rater separation index (equation 7) was estimated and indicates the number of statistically significantly different levels of rater leniency/severity (Myford & Wolfe, 2003). Ideally, the rater separation index will be small, as smaller values indicate fewer statistically distinct levels of rater leniency/severity compared to larger values (Myford & Wolfe, 2004).

The rater reliability of separation (equation 8) was estimated for raters and is an estimate of how reliably raters can be separated along the severity continuum (Myford & Wolfe, 2003). Ideally, the rater reliability of separation will be low, suggesting that raters have similar leniency/severity logits and thus cannot be reliably separated along the ability continuum (Myford & Wolfe, 2003; Myford & Wolfe, 2004).

Additionally, individual raters' logits were evaluated via visual inspection with a Wright map, also known as a vertical ruler or variable map (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2004). The Wright map provided a visual depiction of raters' leniency/severity and the rank-ordering of raters by their leniency/severity logits. Ideally, raters will be clustered around a logit score of 0.0 (i.e. average leniency/severity) on the Wright Map. If raters are dispersed across the logit continuum, it suggests that raters differ in their leniency/severity. Raters who had logit values greater than 0.0, and thus were higher than 0.0 on the Wright map, were considered to be more severe than the average rater. Raters who had logit values less than 0.0, and thus were lower than 0.0 on the Wright map, were considered to be more lenient than the average rater (Bond & Fox, 2015; Eckes, 2015; Linacre, 2017a; Myford & Wolfe, 2004).

As a second visual supplement, confidence intervals were estimated around raters' logit leniency/severity scores using

$$\text{Rater logit} \pm 1.96(SE_{rater}) \quad (14)$$

where $SE_{rater}$ is the standard error of the rater leniency/severity logits (Wolfe, 2004). Because logits are on a continuous scale and will presumably be normally distributed, a critical value of 1.96 was used. Confidence intervals were plotted using SAS Software Version 9.4 (SAS Institute, 2015) to visually determine the extent to which raters differed in their leniency/severity.

In sum, rater leniency/severity were evaluated overall via the fixed-effect chi square, rater separation index, and rater reliability of separation. Each of these aforementioned indices indicate the degree to which raters differ in their leniency/severity. After assessing rater leniency/severity differences globally, individual raters were evaluated visually via the Wright map and confidence intervals. Raters will not be identified by name, but will instead remain as "rater 1," "rater 2," "rater 3," etc. in all results.

**Research question 2: Are there statistically significant rater leniency/severity and ER-WR rubric element interaction effects?** The hybrid MFRM model (equation 3) was modified to include an interaction term in order to evaluate this research question. When evaluating interactions in the MFRM framework, interactions may be tested in an exploratory or confirmatory manner (Eckes, 2015). Exploratory interaction analyses are appropriate when there are no a priori hypotheses about the nature of the interaction. Given that there are no a priori hypotheses regarding the nature of a possible interaction between ER-WR rubric elements and rater leniency/severity, an exploratory interaction

analysis was conducted. One additional interaction term between rater and element was added to equation 3

$$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \varphi_{ij} - \tau_{ik}, \quad (15)$$

where $P_{nijk}$ is the probability of student *n* being rated *k* on element *i* by rater *j*, $P_{nijk-1}$ is the probability of student *n* being rated *k*-1 on element *i* by rater *j*, $\theta_n$ is the ability of student *n*, $\delta_i$ is the difficulty of ER-WR rubric element *i*, $\alpha_j$ is the severity of rater *j*, $\varphi_{ij}$ is the interaction between the severity of rater *j* and ER-WR rubric element *i, and* $\tau_{ik}$ is the difficulty of score level *k* compared to score level *k*-1 on ER-WR rubric element *i* (Eckes, 2015). The interaction parameter, $\varphi_{ij}$, may also be referred to as a bias parameter, as a significant interaction suggests differential functioning of raters across ER-WR rubric elements, or rater bias via ER-WR rubric element (Eckes, 2015).

A two-step calibration procedure was used (Eckes, 2015; Linacre, 2017a; Myford & Wolfe, 2003). In the first calibration, all parameters except $\varphi_{lj}$were estimated. In the second calibration, the parameters from the first calibration were fixed and parameters for $\varphi_{ij}$ were estimated. To evaluate the null hypothesis that rater leniency/severity does not depend on ER-WR rubric elements after controlling for measurement error, the statistical significance of the *t* statistic was evaluated

$$t_{ij} = \frac{\hat{\varphi}_{ij}}{SE_{ij}}, \quad (16)$$

where $\hat{\varphi}_{ij}$ is the estimated parameter for the interaction between the severity of rater *j* and ER-WR rubric element *i*, and $SE_{ij}$ is the standard error of $\hat{\varphi}_{ij}$ (Eckes, 2015). A *t* value for each rater and each ER-WR rubric element was obtained. A statistically significant ($p$<0.05) *t* value suggested a rater differed in his or her leniency/severity

across rubric elements, after controlling for measurement error (Bond & Fox, 2015; Eckes, 2015). To evaluate the ubiquity of rater bias by ER-WR rubric element, the number of $t_{ij}$ indices statistically significant were summed and converted to a percentage that represented the percentage of raters exhibiting leniency/severity bias by ER-WR rubric elements (Eckes, 2005, 2015).

The interaction between rater leniency/severity and ER-WR rubric element was also evaluated visually with a bias diagram. A bias diagram depicts each raters' *t* values for ER-WR rubric elements. Ideally, raters' *t* values for each element will be close together, suggesting minimal bias. If raters' *t* values differ, it suggests that they did not exhibit the same leniency/severity across ER-WR rubric elements (Eckes, 2015). Raters will not be identified by name, but will instead remain as "rater 1," "rater 2," "rater 3," etc. in all results.

**Research question 3: Is there a lack of distinguishability between score levels, suggesting a restriction of range effect? Is this restriction of range suggestive of a central tendency effect?** The hybrid MFRM model (equation 3) was used to evaluate this research question. First, the fixed-effect chi-square was evaluated as a global test of whether students' abilities, as defined by their logit scores, differed. The fixed-effect chi-square (equation 4) was estimated to evaluate the null hypothesis that, after controlling for measurement error, there were no differences in student ability. A statistically significant chi-square ($p$<.05) suggests that at least two students are statistically significantly different in their ability logit scores (Myford & Wolfe, 2004). If students' abilities are indistinguishable (i.e. a non-significant fixed-effect chi square), it suggests that students receive similar scores from raters, revealing a possible restriction of range

effect. Ideally, students will be spread across the logit continuum, suggesting students differ in their ability estimates. Thus, ideally, the chi-square will be significant, suggesting students are spread across the ability continuum, and thus providing evidence that a restriction of range effect may not be present (Myford & Wolfe, 2004).

Next, the student separation ratio, student separation index, and reliability of student separation were estimated with students as the object of measurement. In the student separation ratio (equation 6), the true SD (equation 5) was estimated using the observed standard deviation of the student ability logits and the standard error associated with the student ability logits. Further, the student separation index (equation 7) was estimated. The student separation index indicates the number of statistically significantly different levels of student ability (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003). Ideally, the student separation index will be large, as larger values indicate more statistically distinct levels of student ability compared to smaller values. If distinct levels of student ability are present, it is unlikely that there is a group-level restriction of range effect.

The student reliability of separation (equation 8) was estimated and is an estimate of how reliably students can be separated along the ability continuum. Ideally, the student reliability of separation will be high, suggesting that students vary in their estimated ability logits and thus can be reliably separated along the ability continuum (Myford & Wolfe, 2003).

Additionally, students' ability logits were visually evaluated via a Wright map. Ideally, students will be spread across the logit continuum, suggesting that students differ in their abilities. The average student ability logit was estimated to be -1.40. Thus,

students who had logit values greater than -1.40 were considered to be of higher ability than the average student, and students who had logit values less than -1.40 were considered to be of lower ability than the average student (Myford & Wolfe, 2004).

As an additional supplement to the student separation index and student reliability of separation, frequencies of the scores in each score level for each rubric element were computed. Frequency analyses could have been obtained using a partial credit version of the MFRM model, allowing both the rater and element facets to be partial credit (Myford & Wolfe, 2003). However, for model simplicity, the rater facet was modeled as rating scale, and follow-up frequency analyses were conducted instead. The frequencies were evaluated to determine whether some score levels on the rubric were used more than other score levels. A histogram of the scores was graphed using SAS Software Version 9.4 (SAS Institute, 2015)

Additionally, a cross-tabulation between rater and score level was generated. The frequencies were evaluated to determine whether certain raters used some score levels more than other scores levels. Moreover, frequency and cross-tabulation analyses provided insight regarding the nature of any restriction of range effects. That is, frequency and cross-tabulation analyses provided insight as to whether scores were restricted to 1) the lower or upper ends of the scoring levels, suggesting extreme scoring; or 2) the middle scoring levels, suggesting a central tendency effect. The overall frequency analysis across all raters provides further evidence for a possible restriction of range group-level effect while the cross-tabulation analyses provided further insight into possible restriction of range effects for individual raters.

In sum, central tendency was evaluated overall via the fixed-effect chi square, student separation indices, and student reliability of separation. Each of these aforementioned indices indicates the degree to which students differ in their abilities, thus providing evidence of the extent to which a restriction of range effect may be present. The Wright map provided further visual support for assessing the degree to which students differed in their abilities. As a method for evaluating individual raters' tendencies to exhibit a restriction of range effect, cross-tabulations of raters and score levels were generated. Raters will not be identified by name, but will instead remain as "rater 1," "rater 2," "rater 3," etc. in all results.

**Research question 4: Is there a relationship between raters' knowledge of the 8KQs and rater leniency/severity?** The hybrid MFRM model (equation 3) was used to evaluate this research question. To evaluate the relationship between rater leniency/severity and 8KQ knowledge, raters' leniency/severity logits were correlated with ERIT total scores. A Pearson correlation was evaluated for statistical and practical significance. The correlation was considered statistically significant if $p<0.05$. Cohen's (1992) guidelines for small ($r = 0.1$), medium ($r = 0.3$), and large ($r = 0.5$) effect sizes were used as criteria to interpret the magnitude of the correlation. The relationship between rater leniency/severity and 8KQ knowledge was considered practically significant if it met the guideline for a medium effect size or larger. A visual inspection of the relationship was obtained via a scatterplot of rater leniency/severity and 8KQ knowledge.

In sum, the relationship between rater leniency/severity and 8KQ knowledge was evaluated via a Pearson correlation and visual inspection via a scatterplot of rater leniency/severity by 8KQ knowledge.

# Chapter 4: Results

Four research questions were addressed in this study:

1) Are there statistically significant differences in rater leniency/severity, suggesting a group-level leniency/severity rater effect?
2) Are there statistically significant rater leniency/severity and ER-WR rubric element interaction effects?
3) Is there a lack of distinguishability between score levels, suggesting a restriction of range effect? Is this restriction of range suggestive of a central tendency effect?
4) Is there a relationship between raters' knowledge of the 8KQs and rater leniency/severity?

To address these research questions, an MFRM analysis was conducted on ER-WR scores. In this chapter, findings for each individual research question are presented. However, prior to presenting results, outcomes from assumption testing are provided. To evaluate the formal assumptions of IRT, three MFRM models were estimated: 1) one model in which the MFRM hybrid model (equation 3) was specified for the seven-element structure (i.e. Elements A, B, C, D lower, D upper, E lower, and E upper), 2) one model in which the MFRM hybrid model was specified for five-element structure (i.e. Elements A, B, C, D, and E), and 3) one model in which the MFRM rating scale model (equation 1) was specified for five-element structure. Results from the assumptions testing influenced the choice of the final model and the results interpreted for this study.

## Assumption Testing

**Local independence.** To determine whether the local independence assumption was violated for these data, two analyses were conducted: 1) one analysis in which the MFRM hybrid model (equation 3) was specified for the five-element structure, and 2) one analysis in which the MFRM hybrid model (equation 3) was specified for the seven-element structure. Student reliability of separation indices were compared across each model, and differences less than 0.05 were considered evidence that there was not violation of local independence in these data. Student reliability of separation estimates were nearly identical across both models. Moreover, parameter estimates were compared for Elements A, B, and C across both models, and parameter estimates were similar across both analyses. Correlations between the location and threshold estimates for the five- and seven-element models were greater than 0.99. See Appendix F for a comparison of the reliability and parameter estimates for the two analyses. Similarity of results across models provided evidence that local independence was not violated with these data, so all subsequent analyses and reported results represent the five-element structure.

**Correct model form.** Recall, there were two ways for evaluating fit: 1) overall model fit, and 2) rater fit.

*Overall model fit.* Recall, the element facet was freed to be partial credit due to the different number of possible score options when Elements D and E were split into upper and lower elements. However, because of evidence that the five-element structure adequately met the local independence assumption, there was the option for constraining the element facet to be rating scale. As such, overall model fit was compared for two models: 1) one in which the element facet was treated as partial credit, and 2) one in

which the element facet was treated as rating scale. To evaluate overall model fit, the absolute value of the standardized residuals were evaluated. Residuals greater than |2.0| indicated highly unexpected scores, and data were thought to fit the model well overall if fewer than 5% of the standardized residuals were greater than or equal to |2.0|. For both models, less than 4% of the standardized residuals were greater than or equal to |2.0|. As such, the simpler model where the element facet was constrained to be rating scale was considered adequate. All subsequent analyses and reported results represent the five-element structure with all facets specified as rating scale (equation 1).

*Rater fit.* Recall, to evaluate rater fit, the unweighted mean square ($MS_U$) and weighted mean square ($MS_W$) indices were evaluated for the model in which the five-element structure was maintained and all facets were treated as rating scale. $MS_U$ and $MS_W$ values greater than 2.0 were flagged as indications of major rater misfit. No $MS_U$ nor $MS_W$ values were greater than 2.0. All $MS_U$ values ranged from 0.64 – 1.64, and all $MS_W$ values ranged from 0.67 – 1.76 (see Table 1). Rater three had $MS_U$ and $MS_W$ estimates slightly larger than the preferred range of 0.5 to 1.50, suggesting rater three's ratings were less similar to the model-implied ratings than predicted by the model. However, because rater three's $MS_U$ and $MS_W$ were less than 2.0, rater three's ratings were not considered to be a major distortion in the model. As such, rater fit was considered acceptable for these data.

**Unidimensionality.** Recall that unidimensionality was assessed via conducting a principal components analysis on the standardized residuals. Unidimensionality was assessed using the standardized residuals from the five-element structure with all elements treated as rating scale. Unidimensionality was considered met if there were

fewer than two elements on any secondary contrast. For this study, evidence suggests that ER-WR scores were unidimensional, as all elements loaded on to the first contrast, and eigenvalues were less than 1.0 for each secondary contrast.

**Evaluation of Research Questions**

All research questions were answered via results from the model in which the five-element structure was maintained and all elements were treated as rating scale.

**Research question 1: Are there statistically significant differences in rater leniency/severity, suggesting a group-level leniency/severity rater effect?** First, the fixed-effect chi-square was evaluated to determine whether there were statistically significant differences in rater leniency/severity, after controlling for measurement error. The fixed-effect chi-square was statistically significant ($\chi^2(17) = 2037.3$, $p < .001$), suggesting at least one rater differed significantly in leniency/severity from the other raters. The rater separation index suggested ten statistically distinct levels of rater leniency/severity in this sample of raters. Moreover, the rater reliability of separation was 0.99, suggesting near-perfect separation and rank-ordering of raters' leniency/severity along the leniency/severity continuum.

Because the rater facet was centered at 0.00, leniency/severity estimates of 0.00 indicated average leniency/severity, values less than 0.00 indicated relatively more lenient raters, and values greater than 0.00 indicated relatively more severe raters. For a visual representation of the rank-ordering of raters by their leniency/severity estimates, see the Wright Map (Figure 2). Rater one, rater eleven, rater sixteen, and rater eighteen were of approximately average leniency/severity in this sample of raters (see Table 2). With the exception of rater three and rater seven, raters' leniency/severity were within

one standard deviation (0.68 logits) of average leniency/severity. Rater three and rater seven were both at least one logit more lenient than the average rater. Rater five was the most severe rater, though still within one standard deviation from the average rater. For the extent to which raters differed from one another in their leniency/severity estimates, see Figure 3.

Leniency/severity estimates may be further interpreted in the context of observed scores provided by raters. The average score provided by raters ranged from 0.77 to 2.09 points out of a possible four points on the ER-WR rubric (see Table 2). Based on average observed score estimates, rater four was the most severe and rater three was the most lenient. Note that observed score interpretations do not necessarily align with interpretations from MFRM analyses because the observed scores are not adjusted for the quality of student responses assigned to different raters, as rater five was the most severe according to MFRM estimates, but rater four was the most severe according to the average observed score.

In sum, raters in this sample differed in their leniency/severity. Given that the rater separation index suggested there were ten distinct strata of leniency/severity estimates, it is not surprising that the rater reliability of separation suggested raters could be separated and rank-ordered by their leniency/severity along the leniency/severity continuum. Though no raters were particularly severe in their ratings, two raters were particularly lenient in their ratings.

**Research question 2: Are there statistically significant rater leniency/severity and ER-WR rubric element interaction effects?** To address this research question, a bias parameter was included in the five-element rating scale MFRM model. A bias

parameter was estimated for each rater for each ER-WR rubric element, and the significance of these parameters was evaluated to determine the extent to which raters differed in their leniency/severity across ER-WR rubric elements. Ideally, raters exhibit similar, and average, leniency/severity across ER-WR rubric elements, resulting in non-significant bias parameters across rubric elements. The bias diagram (see Figure 4) was used as a visual depiction of raters' interaction parameters (see Table 3). Bias parameters greater than 0 indicated raters' scores were more lenient than expected by the model, and bias parameters less than 0 indicated raters' scores were more severe than expected by the model (Eckes, 2015). If the bias parameters differed across ER-WR rubric elements for individual raters, it suggested that raters' leniency/severity differed depending on ER-WR rubric element. Many raters exhibited differential leniency/severity across ER-WR rubric elements. There were 90 total bias parameters (18 raters*5 elements). If bias were not present in scores, we might expect approximately four bias parameters to be significant by chance alone. However, of the 90 total bias parameters, 33.33% were statistically significant, suggesting that, after controlling for measurement error, raters differed in their leniency/severity across rubric elements about one-third of the time. Thus, many more raters exhibited bias across the ER-WR rubric elements than would be expected by chance. Interestingly, of the 30 statistically significant bias parameters, 11 were for Element A.

Rater seven, rater eight, rater twelve, rater thirteen, rater fourteen, and rater eighteen were relatively consistent in their leniency/severity across elements and did not exhibit statistically significant bias across any elements. Rater three differed extensively in leniency/severity across rubric elements and exhibited statistically significant bias on

all elements. Compared to other elements, rater three was the most severe on Element E and the most lenient on Element B. Several raters (i.e. rater two, rater six, rater nine, rater eleven, and rater seventeen) were similar in their leniency/severity across Elements B, C, D, and E, but differed in their leniency/severity for Element A compared to the other elements. Rater two, rater eleven, and rater seventeen were much more lenient on Element A compared to the other elements and rater six and rater nine were more severe on Element A compared to the other elements.

In sum, several raters' observed scores were more lenient or more severe than expected by the model, resulting in significant bias parameters. Whether the raters' observed scores were more lenient or more severe than expected differed across ER-WR rubric elements for many raters, suggesting differential leniency/severity across ER-WR rubric elements for several raters in this sample. Of the five ER-WR rubric elements, Element A in particular appeared to be problematic for raters, as over one-third of the significant bias parameter estimates were for Element A.

**Research question 3: Is there a lack of distinguishability between score levels, suggesting a restriction of range effect? Is this restriction of range effect suggestive of a central tendency effect?** The fixed-effect chi-square was statistically significant ($\chi^2(464) = 5723.7$, $p < .001$), suggesting that, after controlling for measurement error, students differed significantly in their abilities. The student separation index suggested there were 2.73 distinct groups of students. A student reliability of separation of 0.88 suggests that students' abilities are rank-ordered fairly consistently (across raters and rubric elements) along the ability continuum. For a visual depiction of the separation of

students' abilities along the ability continuum, see Figure 2. The student facet was not centered and thus the average student ability was estimated to be -1.40 logits.

Observed score frequencies provided by raters in each score category indicated approximately two-thirds of scores provided by raters were between 0 – 1, and few scores were provided in the 3.5 and 4 score categories (see Table 4). Moreover, this same pattern was present in many raters' scores. Thus, it appears that a restriction of range effect was present for many raters, with the restriction limited to the lowest three score categories. Note that though these results suggest a restriction of range effect for raters, it could be the case that students were of low abilities, resulting in an accurate restriction of raters' scores to the lowest score categories.

Ideally, raters provide similar scores to the same student essays, such as was the case with rater seventeen and rater eighteen (see Table 4). Most rater pairs were fairly similar in the scores they provided to student essays. However, rater three and rater four greatly differed in the scores they provided to student essays. Rater three did not exhibit a restriction of range effect, as rater three provided scores across the spectrum of the rubric. On the other hand, rater four exhibited restriction of range, as rater four provided scores primarily in the lower score categories of 0, 0.5, and 1.

In sum, the student reliability of separation and student separation index suggested that a restriction of range effect was not present in these data. However, observed score frequencies suggested that most raters restricted their scores to the lowest score categories of the ER-WR rubric. Together, this information suggests that, though raters exhibited some restriction of range, as was evidenced by review of the observed

score frequencies, students' scores still differed enough to allow for students' abilities to be adequately separated and rank-ordered along the ability continuum.

**Research question 4: Is there a relationship between raters' knowledge of the 8KQs and rater leniency/severity?** Total scores from the Ethical Reasoning Identification Test (ERIT) and raters' leniency/severity estimates were correlated to determine whether there was a relationship between raters' knowledge of the 8KQs and their leniency/severity. The Pearson correlation representing the relationship between raters' knowledge of the 8KQs and rater leniency/severity was non-significant ($r = 0.37$, $p = 0.14$). By Cohen's (1992) effect size guidelines, the relationship between raters' knowledge of the 8KQs and their leniency/severity is a medium effect. However, because of the small sample size of the rater pool, the magnitude of the relationship between raters' 8KQ knowledge and leniency/severity should be interpreted cautiously. Moreover, review of the scatterplot between raters' ERIT scores and leniency/severity logits (see Figure 5) illuminates that the correlation between raters' ERIT scores and leniency/severity was inflated by rater seven. If rater seven was removed from the analysis, the relationship between raters' ERIT scores and leniency/severity dropped to 0.08, a negligible effect size. Thus, the moderate effect size found in this study was likely due to rater seven, who appeared to be an outlier.

Additionally, when evaluating the scatterplot between raters' ERIT scores and leniency/severity logits, it is difficult to distinguish a discernible pattern. Rater eleven, who achieved a perfect score on the ERIT and had the highest ERIT total score, was of average leniency/severity. However, rater nine and rater sixteen were also of average leniency/severity, yet achieved two of the three lowest ERIT total scores. Rater three,

who was the most lenient rater, achieved an ERIT total score of 34, which was similar to raters five and rater fourteen who were the two most severe raters.

In sum, a strong relationship between raters' 8KQ knowledge and leniency/severity was not discernible, especially if rater seven was not included in the analysis. Though the overall relationship was positive on average across all raters, knowing raters' 8KQ knowledge did not necessarily provide useful information about raters' leniency/severity, or vice versa.

# Chapter 5: Discussion

This study was designed to evaluate the extent to which rater effects influence students' ER-WR scores. Specifically, I examined differences in raters' leniency/severity, the extent to which raters' leniency/severity differed across ER-WR rubric elements, and whether raters exhibited restriction of range through their scores. Though these findings provide some context regarding *how* raters behave, the MFRM analyses do not provide information regarding *why* raters' leniency/severity may differ. Thus, in a preliminary effort to explain why raters may differ in their leniency/severity, I evaluated whether raters' knowledge of the 8KQs was related to their leniency/severity. A discussion of each research question and general findings is presented below. Implications of the results and directions for future research are discussed.

## Research question 1: Are there statistically significant differences in rater leniency/severity, suggesting a group-level leniency/severity rater effect?

In this study, raters differed in their leniency/severity. At least ten distinct strata of raters were identified from the MFRM analysis. Ideally, raters will not be separable into distinct strata, suggesting raters are interchangeable. If raters are interchangeable, then the score students receive do not depend upon the rater who scores their essay. Findings from this study are not ideal and suggest that students' unadjusted scores may depend upon which rater evaluated their essay.

Interestingly, no raters were especially severe in their ratings. However, rater three was particularly lenient compared to other raters, receiving a leniency/severity logit 1.5 standard deviations below the average rater. Additionally, rater three was the only rater to have an average observed score across all student essays above two points on the

zero to four ER-WR rubric scale, double the observed sum score of several other raters. For another perspective, rater three's scores may be compared to rater four's scores, as raters three and four were rating partners and rated the same essays. Despite rating the same essays, rater three provided an average observed score that was more than twice the average observed score provided by rater four. Clearly, there are large discrepancies between rater three and rater four's scores, raising questions about the meaning of students' scores for their essays.

Often, essay scores are averaged across rater partners before providing descriptive information about students' scores to stakeholders. Rater three and rater four rated so differently from one another that their average score likely fails to accurately represent students' abilities. For example, consider for Element A that rater three provided a score of three to an essay, and rater four provided a score of one to the same essay, resulting in an average score of two on Element A for that student. According to the ER-WR rubric, stakeholders could interpret a score of two as a representation that the student explicitly referenced ethical decision options, but did so in a disorganized manner. However, given the discrepancies between rater three and rater four's scores, is the student's ability truly indicative of a score of two? If a score of three is accurate, then the student explicitly referenced ethical decision options and provided a clear and organized account of the ethical decision options. If a score of one is accurate, then the student did not provide an explicit reference to ethical decision options. Thus, the interpretations stakeholders make vary depending on whether they interpret rater three's score, rater four's score, or the average score. Ideally, stakeholders use results to create targeted interventions that improve students' skills and abilities. If discrepant scores are averaged and do not

represent students' true ethical reasoning abilities, it is challenging for stakeholders to use the results in meaningful ways. In short, when raters are as discrepant as raters three and four, it is difficult to know which score reflects a student's ethical reasoning abilities, creating implications for score interpretation and educational programming.

**Research question 2: Are there statistically significant rater leniency/severity and ER-WR rubric element interaction effects?**

In this study, several raters differed in their leniency/severity across ER-WR rubric elements, thereby exhibiting bias. Some raters (e.g. rater seven, rater eight, rater twelve, rater thirteen, rater fourteen, and rater eighteen) did not exhibit significant bias across any ER-WR rubric elements, suggesting consistent application of the rubric criteria across all rubric elements. However, other raters varied drastically in their leniency/severity across ER-WR rubric elements, suggesting inconsistent application of the rubric criteria across all rubric elements. Of any rater, rater three differed the most in leniency/severity across elements and exhibited significant bias on all rubric elements.

Interestingly, raters' scores were the most biased for Element A compared to other elements. Historically, raters anecdotally suggest that Element A is challenging to rate, and given that nearly half of the significant bias parameters across raters were for Element A, empirical results from this study support anecdotes from previous raters. Several raters (e.g. rater six, rater nine, rater seventeen) were relatively consistent in their leniency/severity across Elements B, C, D, and E, but exhibited significant bias for Element A. Interestingly, bias did not always manifest as either leniency or severity. That is, raters differed in whether they were significantly more severe or more lenient on Element A compared to other elements. When evaluating the ER-WR rubric scoring

criteria, Element A appears to be one of the more subjective elements, providing context for why differential bias was observed for Element A relative to the other elements.

During the rating session for the current study, many raters asked clarification questions about Element A and sought advice regarding whether students actually presented an ethical situation. Moreover, several raters indicated that essays did not have ethical situations when they did actually have an ethical situation present. If raters indicated that an ethical situation was not present, then they typically scored the essay very low, often assigning scores of 0 to all elements, which could be contributing to the leniency/severity bias found for Element A in this study. If raters are unable to adequately identify the content that needs to be rated, their leniency/severity is likely to differ for the element containing the content that is challenging to identify. In such a case, additional training may be provided to raters. In the context of the Madison Collaborative, staff may consider adding a module during which raters have the opportunity to identify ethical situations.

Though many raters exhibited differential leniency/severity for Element A, many raters exhibited consistent leniency/severity on Element B and Element C. That is, when ratings were lenient on Element B, they were also lenient on Element C, and when ratings were severe on Element B, they were also severe on Element C. Consistency across Element B and Element C is not surprising, as Element B and Element C each require raters to count KQs and provide a score based on the number of KQs students reference and provide a rationale for, respectively. Because these elements involve a simple counting of KQs, it makes sense that raters exhibited minimal bias across these elements.

Note that, though it is ideal raters exhibited minimal bias on Element B and Element C, it does not necessarily mean that raters apply the 8KQs correctly or have adequate knowledge of the 8KQs. That is, just because raters exhibit similar leniency/severity on Element B and Element C, they may not correctly identify KQs in students' essays. Rather, raters may consistently misidentify KQs in students' essays, producing consistent, but incorrect, scores. Future researchers may focus on whether raters identify the correct KQs in students' essays. Though raters' leniency/severity was not as consistent across Element D and Element E as for Element B and Element C, most raters did not exhibit significant bias on Element D or Element E.

**Research question 3: Is there a lack of distinguishability between score levels, suggesting a restriction of range effect? Is this restriction of range effect suggestive of a central tendency effect?**

Restriction of range may occur at any score level, resulting in central tendency or extreme rating effects. In this study, most raters appeared to exhibit a restriction of range effect, with scores restricted to the lowest three score levels. This result is not surprising, given that students historically score less than a 2.0 on average across all elements. Thus, the restriction of range may be warranted, given students' abilities.

Though the restriction of range is not particularly concerning, it is concerning that some raters provided scores in certain score categories more often than other raters. With the exception of rater three, all raters had positively skewed ratings, with more ratings at the low end of the score levels and few ratings at the high end of the score levels. On the other hand, rater three, the most lenient rater, provided more scores of four than any other score and used all score categories with similar frequency. Such a score distribution is

unlikely for two reasons. First, historical trends indicate students typically do not score above a two on average across all ER-WR rubric elements, making it unlikely that so many high scores were warranted, given students' abilities. Second, rater four, who rated the same essays as rater three, did not provide a similar number of scores across the score levels. Though we cannot be sure that rater four's ratings are accurate, rater four provided few scores above a score of two. Thus, in comparison to rater four and average scores from previous years' rating sessions, rater three appeared unnecessarily lenient.

Rater five, rater six, rater nine, and rater ten provided more scores of zero than any other raters. These raters each provided more than 100 ratings of zero across all of their essays, whereas many other raters provided less than 70 ratings of zero. Although not indicated particularly severe from the MFRM analysis, it may be possible that these raters are relatively severe raters. Or, it could be possible that these raters happened to receive essays of less quality compared to other raters. This second explanation is plausible considering that rater five and rater six were a rater pair and rater nine and rater ten were a rater pair. Because each rater pair rated the same essays, and provided a similar number of zeros to student essays, it could be the case that they received essays of lower quality simply by chance. Their comparative influx of zero ratings compared to other raters would be more concerning if their rating partners did not provide a similar frequency of zeros to student essays. As previously discussed, though rater effects were identified via MFRM analysis, the analysis does not necessarily provide information regarding *why* raters exhibit rater effects. As such, results are next presented for research question four, which was included in this study to identify reasons why raters may exhibit rater effects.

**Research question 4: Is there a relationship between raters' knowledge of the 8KQs and rater leniency/severity?**

In this study, raters' knowledge of the 8KQs was unrelated to their leniency/severity. These findings suggest that knowing raters' 8KQ knowledge does not necessarily inform how lenient/severe raters will be, or vice versa. This research question was examined as a potential method to diagnose why raters exhibit differential leniency/severity. However, results did not provide an explanation for why raters differed in their leniency/severity. Rater three, the most lenient rater, scored near the average Ethical Reasoning Identification Test (ERIT) score. The two most severe raters also scored near the average ERIT score. Considering the most lenient and most severe raters scored similarly on the ERIT, raters' knowledge of the 8KQs did not provide useful information regarding raters' leniency/severity.

In this study, 8KQ knowledge was measured via the ERIT. There are concerns regarding the validity of ERIT scores as representations of 8KQ knowledge. Though the ERIT was the best readily available option for assessing raters' 8KQ knowledge in this study, it is not a perfect assessment and may not be the best measure of raters' knowledge of the 8KQs. Stakeholders plan to revise the assessment to address concerns surrounding some of the questions on the ERIT. It may be beneficial to replicate this research question in a future study with the revised ERIT.

## General Discussion

**Benefits of MFRM.** To obtain a general picture of rater leniency/severity without the use of MFRM, observed score averages or observed sum scores such as those in Table 2 may be obtained. When ranked by observed average score, rater leniency/severity

rank-ordering just slightly differs from the rank-ordering of raters produced by the MFRM analysis. The similarity of rank-ordering of raters by observed scores and leniency/severity estimates likely occurred due to the random assignment of essays to raters. That is, because raters scored essays of approximately equal quality, their rank-ordering was the same whether evaluating observed scores or MFRM estimated leniency/severity. Consequently, similar interpretations may be drawn about raters' leniency/severity regardless of whether observed scores are used or whether an MFRM analysis is conducted. The advantage of MFRM lies in the other information gathered from the analysis.

For example, rater infit and outfit provide information about the extent to which raters' scores were expected, given the specified MFRM model. Infit/outfit estimates specifically may be used in conjunction with raters' leniency/severity estimates to provide additional context for whether researchers choose to remove ratings from particular raters who appear problematic. For this study, many raters had values near 1.00, suggesting they provided scores similar to those expected by the model. Rater three approached the upper bounds of acceptable infit/outfit estimates, suggesting rater three assigned scores that were unexpected, given the model. Given rater three's extreme leniency compared to other raters, and rater three's large infit/outfit values, the Madison Collaborative may consider whether rater three's ratings are too different from those expected by the model to include in the final ER-WR essay results presented to stakeholders.

Additionally, MFRM analyses provide researchers with "fair average" scores. The fair average scores represent the score a student would have received, had they been

evaluated by a rater of average leniency/severity (Eckes, 2015). If rater leniency/severity is a concern, students' observed scores may be replaced by fair average scores to remove the leniency/severity effect from the scores. The estimation of fair average scores is an advantage of MFRM.

An additional benefit of MFRM above other techniques such as g-theory is the ability to identify individual raters who may be problematic. G-theory is a common technique used to evaluate reliability for performance assessment scores, and variance components from g-theory may be useful for determining whether raters differ in their leniency/severity. In this study, differences in rater leniency/severity accounted for approximately 8% of variability in student scores. Though g-theory results suggest raters differ in their leniency/severity, g-theory results cannot be used to identify individual raters who may be lenient or severe. G-theory has benefits as a group-level technique, but if researchers desire diagnostic information about individual raters, g-theory is not a useful technique. When using MFRM, researchers obtain individual rater information, which is beneficial if the goal is to identify raters for additional training opportunities or adjust scores after analysis.

**Limitations of MFRM.** MFRM is a large sample technique. Currently, the Madison Collaborative collects several hundred student essays each year, making MFRM feasible. However, if researchers do not have the resources to collect and score a large sample of essays, MFRM may not be a feasible method for evaluating scores. Additionally, MFRM analyses require up-front planning, as raters must each rate a common sub-sample of student essays. It is necessary to rate common sub-sample of student essays in order to create connected subsets of raters that allow for rater estimates

to be compared to one another. A benefit of MFRM is that raters and all other facets are placed on a common logit scale. However, this benefit may only be actualized if raters' scores are connected via a common sub-sample of student essays.

When essays are rated in such a way that there is a common sub-sample of student essays all raters have rated, MFRM is a normative technique. That is, MFRM allows for raters to be compared on a common scale. Because of the normative nature of MFRM, rater three was found to be more lenient than the average raters in this sample of raters. Moreover, rater three applied the score levels differently than the other raters. Given this information, it appears as though rater three is a "poor" rater and many other raters are "good" raters in this sample of raters. However, the normative nature of MFRM raises the question of whether rater three is actually a lenient rater, or if rater three is actually rating appropriately based on rubric criteria. MFRM analyses alone cannot answer this question; additional information is needed to supplement MFRM results. In this study, we have empirical historical information about students' scores that is helpful for interpreting MFRM results. As previously discussed, this empirical information from prior years suggests rater three is likely overly lenient and is not rating in accordance with the ER-WR rubric criteria. However, researchers may not always be in a position where they are privy to previous research or scores. If researchers do not have additional information to assist with interpreting information about the ratings, they must always keep in mind the limitation of the normative nature of MFRM.

Recall, MFRM analyses produced rater leniency/severity estimates that rank-ordered raters similarly to their observed scores. That is, we come to similar conclusions regarding raters' leniency/severity whether we rank-order raters by their observed scores

or their MFRM leniency/severity estimates. As such, if a goal is simply to identify which raters are more severe or more lenient compared to other raters, an MFRM analysis may not be necessary, assuming essays are randomly assigned and raters are not systematically assigned essays of different quality. Conducting MFRM analyses requires specialized software (e.g. FACETS; Linacre, 2017b) and knowledge of measurement theory to conduct the analysis and evaluate results. Thus, MFRM may not produce substantial information about raters' leniency/severity that cannot already be obtained by evaluating raters' observed scores. Researchers must consider whether the information and benefits provided by MFRM (e.g. rater infit/outfit values, fair average scores, ability to evaluate all facets on a common logit scale) are worth the purchase of additional software and challenges associated with conducting the analysis and interpreting results.

**ER-WR Rubric "Special Notes".** Prior to data analysis, it was expected that scores would violate the local independence assumption necessary for MFRM. This was expected due to the "special notes" on the ER-WR rubric that restrict raters' scores on Elements D and E, depending on the scores assigned to Elements C and D. If local independence were violated, reliability was expected to be inflated for the five-element model and the parameter estimates would differ for Elements A, B, and C across the five- and seven-element models. However, reliability and parameter estimates were nearly identical regardless of whether the five- or seven-element model was specified (See Appendix F). As such, data did not appear to violate local independence. Data may not have violated local independence due to the fact that students tend to receive low scores, thus voiding the need for the special notes and eliminating score dependencies created by the special notes. In the future, researchers may consider evaluating the extent to which

data are dependent in a sample of essays that span the ability spectrum. If raters have the opportunity to provide higher scores, the special notes may become relevant, thus introducing dependencies into ER-WR scores.

If the special notes are used and ER-WR scores do violate local independence, it is worth considering additional avenues through which the data could be modeled. For this study, I intended to model the special notes by creating upper and lower elements for Element D and Element E. Though this model should have been adequate for accounting for dependencies in the data, it required running the element facet as partial credit to handle the differences in score options between the upper and lower elements. Though it is easy to specify an element as partial credit in FACETS (Linacre, 2017a), doing so creates additional estimation needs and requires a larger sample. Additionally, results from partial credit facets may be challenging to interpret, potentially making the results inaccessible to those without a measurement background. As such, it may be beneficial for researchers to consider other options for accounting for data, such as testlet models or Bayesian estimation procedures.

In addition to modeling challenges, the special notes anecdotally pose a challenge to some raters. Specifically, some raters report disagreement with the special notes or do not understand the purpose of the special notes. In effect, some raters have intentionally chosen not to use the special notes, thereby failing to adhere to the rubric scoring criteria. Moreover, raters may forget to use the special notes. In this study, raters did not adhere to the special notes for eight essays, and these essays were removed from the analysis. Though lack of adherence to the special notes was only an issue for a few essays, failure to adhere to the special notes could drastically alter the average score students receive

from both raters, and negatively influence inter-rater reliability. Perhaps additional emphasis could be placed on the special notes during training. Or, the Madison Collaborative may explore additional methods for reminding raters about the special notes, such as verbal reminders during rating, or electronic reminders that are intentionally placed during the rating session.

## Implications

**Score adjustment.** Results from this study provide evidence that raters were not interchangeable, raising concerns about the meaning of ER-WR scores. The next question is how to handle scores from raters who stand out as "poor" raters, such as rater three in this study. One option is to remove rater three's scores. Completely removing rater three's scores is not ideal, as rater four is then the only rater who provided scores to the student essays assigned to this pair of raters. Rater four is a relatively severe rater who differed in his or her leniency/severity across ER-WR rubric elements, so scores from rater four also may not be the best representations of students' ethical reasoning abilities.

A preferable option is to use the fair average scores from the MFRM analysis to adjust students' scores. As discussed, the fair average scores are the scores students would have received if evaluated by a rater of average leniency/severity. The Madison Collaborative could use the fair average scores to adjust students' scores. An advantage of this option is that the rater leniency/severity effect is essentially removed from scores. A disadvantage of this approach is that it may be challenging to explain to stakeholders that scores were adjusted based on a statistical model. Stakeholders may not buy into the idea that scores produced from a statistical model are preferable to the observed scores provided by raters.

Moreover, adjusting students' scores may have implications for the inferences made from scores. Recall, a university strategic plan goal is that students will achieve an average score of two on the ER-WR rubric by 2020. As such, the Madison Collaborative annually reports the percent of students who meet the benchmark of an average score of two on the rubric. With the observed scores in this study, 5.8% of students met the benchmark. However, with the fair average scores, 9.5% of students met the benchmark. For the distributions of scores, see Figure 6. Though the distributions of scores did not change dramatically, the distribution of fair average scores did shift slightly compared to the distribution of observed scores, resulting in more students meeting the benchmark of two when the fair average scores are used. Whether the Madison Collaborative chooses to maintain the observed scores or replace them with the fair average scores has implications for the inferences made regarding the strategic plan goal.

Additionally, if fair average scores are used, scores are no longer comparable to previous years' scores. If longitudinal data is important for research, researchers must consider if and when to begin adjusting students' scores. If scores are adjusted to account for rater leniency/severity, the average score across all essays could potentially differ in a meaningful way, making it appear as though students' abilities for a cohort changed when they, in reality, did not. If the Madison Collaborative intends to present trends of scores across time in relation to the benchmark of two, they should not adjust scores until after 2020.

**Rater training.** As discussed, an advantage of MFRM is that it can be used to identify individual raters who may need additional training. Specifically, researchers may monitor raters during the rating session and use ratings from day one to inform training

on day two. Because rating typically occurs over several days, an MFRM analysis could be conducted on ratings gathered from the first day to identify raters who are rating particularly leniently or severely. Prior to the second day, raters could then receive individualized training to direct them toward an average leniency/severity level. A disadvantage of this method is that the training could be resource-intensive, particularly if individualized training is provided. Moreover, raters may be intimidated by the idea that their scores are being evaluated, thereby introducing unexpected central tendency effects in order to avoid being identified as a particularly lenient or severe raters. On the other hand, using MFRM to identify individual raters could also provide raters with the attention that they need to ensure that they closely follow the rubric scoring criteria, thereby generating scores that are adequate representations of students' abilities.

**Conclusion**

Previous research has documented the challenges related to performance assessments, particularly regarding the manner in which raters provide scores to student work. This study provides additional evidence for the challenges associated with performance assessments, particularly in the assessment of ethical reasoning. Specifically, scores were evaluated for rater effects, and results suggest meaningful differences in rater leniency/severity. Moreover, results suggest a restriction of range effect is present, though restriction of range may be warranted due to restrictions in students' abilities. The evaluated rater effects are relevant not only to this study, but to performance assessments in general.

The majority of current rater effects studies provide evidence that raters' leniency/severity differs, but few published studies explore why raters' leniency/severity

differs. Unique to this study was the attempt to determine why raters differed in their leniency/severity in the context of ethical reasoning essay assessment. Unfortunately, meaningful results were not found, as no relationship between raters' knowledge of ethical reasoning concepts and their leniency/severity was observed for this study. In future research, it is recommended that researchers focus on understanding why raters' leniency/severity differs. It is only through exploring the *why* behind rater effects that researchers and assessment professionals may effectively mitigate rater effects in performance assessment scores.

In addition to evaluating rater effects in the context of ethical reasoning, this study provides an example of how many-facets Rasch measurement (MFRM) models may be used to evaluate performance assessment scores for rater effects. Though MFRM provides benefits above traditional performance assessment methods (e.g. generalizability theory, inter-rater reliability indices), MFRM also creates logistical challenges. Most notably, MFRM analyses require up-front planning, large sample sizes, and knowledge of measurement principles. This study may be useful for researchers and assessment professionals to determine whether the benefits of MFRM are worth the additional challenges associated with the technique.

## References

American Association of Colleges and Universities. (2015). *An introduction to LEAP: Liberal education & America's promise, excellence for everyone as a nation goes to college.* Retrieved from http://www.aacu.org/sites/default/files/files/LEAP/IntroToLEAP2015.pdf

American Association of Colleges and Universities. (2017). *On solid ground: Value report 2017.* Retrieved from http://www.aacu.org/sites/default/files/files/FINALFORPUBLICATION RELEASEONSOLIDGROUND.pdf

American Association of Colleges and Universities. (n.d.). *Value.* Retrieved from https://www.aacu.org/value

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Arum, R. & Roksa, J. (2011). *Academically adrift.* Chicago, Il: The University of Chicago Press.

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly, 9,* 225-248.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences.* New York: NY: Guilford Publications.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12,* 86-107.

Bashkov, B., Smith, K. L., Fulcher, K. H., & Sanchez, E. H. (2014). *Madison Collaborative Annual Assessment Report #1*.

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

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

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of Applied Psychology, 60,* (556-560).

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer-Verlag.

Cizek, G. J. (1991a). Innovation or enervation: Performance assessment in perspective. *The Phi Delta Kappan, 73,* 150-153.

Cizek, G. J. (1991b). Confusion effusion: A rejoinder to Wiggins. *The Phi Delta Kappan, 72,* 695-699.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37,* 163-178.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Cronbach, L. J. (1990). *Essentials of psychological testing.* New York, NY: Harper & Row, Publishers, Inc.

Darling-Hammond, L. (2014). Introduction: The rationale and context for performance assessment. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 1-14). San Francisco, CA: Jossey-Bass.

DeCotiis, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance, 19,* 247-266.

DeMars, C. E. (2010). *Item response theory*. New York, NY: Oxford University Press, Inc.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, 5,* 7-29.

Dimensionality: contrasts and variances. (n.d.) Retrieved January 20, 2018 from http://winsteps.com/winman/principalcomponents.htm

Downing, S. M. (2006). Selected response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.). *Handbook of test development* (pp. 287-302). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

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

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25,* 155-185.

Eckes, T. (2009). Many-facet rasch measurement. In S. Takala (Ed.). *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching,*

*assessment* (Section H.). Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Frankfurt am Main, Peter Lang GmbH.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work?. *Language Assessment Quarterly, 2,* 175-196.

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

Engelhard, G. Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted rasch model. *Journal of Educational Measurement, 31,* 93-112.

Engelhard, G. Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33,* 56-70.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal and T. Haladyna (Eds.), *Large-scale Assessment Programs for ALL Students: Development, Implementation, and Analysis,* (pp. 261-287). Mahwah, NJ: Erlbaum.

Fisicaro, S. A. & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14*(4), 419-429.

Good, M. R. (2015). *Improving student learning in higher education: A mixed methods study* (Unpublished doctoral dissertation). James Madison University.

Gronlund, N. E. (2003). *Assessment of student achievement.* Boston, MA: Pearson Education, Inc.

Guilford, J. P. (1954). *Psychometric methods.* New York, NY: McGraw-Hill Book Company, Inc.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Gaithersburg, MD: Advanced Analytics, LLC.

Haertel, E. H. (1999). Performance assessment and educational reform. *Phi Delta Kappan, 80,* 662-666.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (pp. 65-110). New York: American Council on Education & Praeger.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items.* Mahwah, NJ: Lawrence Erlbaum Associates.

Han, C. (2015). Investigating rater severity/leniency in interpreter performance testing: A multifaceted Rasch measurement approach. *Interpreting, 17,* 255-283.

Hardy, R. (1995). Examining the costs of performance assessment. *Applied Measurement in Education, 8,* 121-134.

Hardy, R. (1996). Performance assessment: Examining the costs. In M. B. Kane & R. Mitchell (Eds.) *Implementing performance assessment: promises, problems, and challenges* (pp. 107-118). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Hart Research Associates (2015). Falling short? College learning and career success: Selected findings from online surveys of employers and college students

conducted on behalf of the Association of American Colleges & Universities. Retrieved from https://www. aacu.org/sites/default/files/files/LEAP/2015employerstudentsurvey.pdf.

Hathcoat, J. D., Penn, J. D., Barnes, L. L. B. Comer, J. C. (2016). A second dystopia in education: Validity issues in authentic assessment practices, *Research in Higher Education, 57,* 892-912.

Holzman, M. A. (2016). *Exploring the validity of ethical reasoning and writing (ER-WR) essay scores through raters' cognitive scoring processes.* Unpublished manuscript.

Holzman, M. A., Ames, A. J., & Pyburn, L. (2017). *Madison Collaborative Annual Assessment Report #4.*

Humphry, S. M. & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43,* 253-263.

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

James Madison University (2013). *The madison collaborative: Ethical reasoning in action.* Retrieved from: https://www.jmu.edu/files/qep-proposal.pdf

JMU Office of Institutional Research (2017). James Madison University strategic plan performance measures. Retrieved from: https://www.jmu.edu/jmuplans/_docs/ StrategicPlanMeasures.pdf

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

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

Khattri, N., Reeve, A. L., & Kane, M. B. (1998). *Principles and practices of performance assessment,* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Kuh, G. D., Ikenberry, S. O., Jankowski, N. A., Cain, T. R., Ewell, P. T., Hutchings, P., Kinzie, J. (2015). *Using evidence of student learning to improve higher education.* San Francisco, CA: Jossey-Bass.

Landy, F. J. & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87,* 72-107.

Lane, S. (2014). Performance assessment: The state of the art. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 133-184). San Francisco, CA: Jossey-Bass.

Lane, S. & Stone, C.A. (2006). Performance Assessments. In B. Brennan (Ed.), *Educational measurement* (pp. 387-432). New York: American Council on Education & Praeger.

Leckie, G. & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48,* 399-418.

Lenz, B. Wells, J. & Kingston, S. (1991). *Transforming schools using project-based learning, performance assessment, and common core standards.* San Francisco, CA: Jossey-Bass.

Li, Y., Li, S., & Wang, L. (2010). *Application of a general polytomous testlet model to the reading section of a large-scale English language assessment.* Princeton, NJ: Educational Testing Service.

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

Linacre, J. M. (2003). Rasch power analysis: Size vs. significance: Infit and outfit mean-square and standardized chi-square fit statistic. *Rasch Measurement Transactions, 17*(1), 918.

Linacre, J. M. (2017a). A user's guide to Facets Rasch-Model computer programs. Winsteps.com.

Linacre, J. M. (2017b) Facets computer program for many-facet Rasch measurement, version 3.80.0. Beaverton, Oregon: Winsteps.com

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1-16.

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

Long, H. & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity, 15,* 13-25.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters?. *Language Testing, 19,* 246-276.

Lumley, T. & McNamara, T. F. (August, 1993). *Rater characteristics and rater bias: Implications for training.* Paper presented at the Language Testing Research Colloquium. Cambridge, England, United Kingdom.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3,* 331-345.

Madaus, G. F. & Kellaghan, T. (1993). The British experience with 'authentic' testing. *Phi Delta Kappan, 74*, 458-469.

Marais, I. & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 200-215.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69,* 147-156.

McNamara, T. (1996). *Measuring second language performance,* Harlow, England: Pearson Education Limited.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 11-28). Washington D.C.: National Center for Educational Statistics.

Moskal, B. M. (2000). Scoring rubrics: What, when, and how?. *Practical Assessment, Research, & Evaluation, 7*(3). Available online: http://pareonline.net/getvn.asp?v=7&n=3.

Moskal, B. M. & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, & Evaluation, 7*(10).

Muraki, E., Hombo, C. M., Lee, Y. W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24,* 325-337.

Murphy, K. R. & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74,* 619-624.

Murphy, K. R. & Cleveland, J. (1991). *Performance appraisal: An organizational perspective,* Boston, MA: Allyn and Bacon.

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.

Nisbett, R. E. & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments, *Journal of Personality and Social Psychology, 35,* 250-256.

Picus, L. O., Adamson, F., Montague, W. & Owens, M. (2010). *A new conceptual framework for analyzing the costs of performance assessment.* Stanford, CA: Stanford Center for Opportunity Policy in Education.

Pinot de Moira, A., Massey, C., Baird, J. A., & Morrissy, M. (2002). Marking consistency over time. *Research in Education, 67,* 79-87.

Powers, D. & Kubota, M. (1998). Qualifying essay readers for an online scoring network. (Research Report RR-96-20). Princeton, NJ: Educational Testing Service.

Resnick, D. P. & Resnick, L. B. (1996). Performance assessment and the multiple functions of educational measurement. In M. B. Kane & R. Mitchell (Eds.) *Implementing performance assessment: promises, problems, and challenges* (pp. 23-38). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Rezaei, A. R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 5*(1), 18-39.

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

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

SAS Institute (2015). SAS software version 9.4. Cary, North Carolina: SAS Institute, Inc.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25,* 465-493.

Schafer, W. D., Gagné, P., & Lissitz, R. W. (2005). Resistance to confounding style and content in scoring constructed-response items. *Educational Measurement: Issues and Practice, 24*(2), 22-28.

Schmeiser, C. B. & Welch, C. J. (2006). Test development. In B. Brennan (Ed.), *Educational Measurement* (pp. 307-354). New York: American Council on Education & Praeger.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: expert readers versus lay readers. *Language Testing, 14,* 157-184.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85,* 956-970.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215-232.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Smith, E. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement, 6*(2), 147-163.

Smith, K. L. (2017). *Integrating Implementation Fidelity and Learning Improvement to Enhance Students' Ethical Reasoning Abilities* (Unpublished Doctoral Dissertation). James Madison University.

Smith, K. L., Fulcher, K. H., & Pyburn, L. (2015). *Madison Collaborative Annual Assessment Report #2*.

Smith, K. L., Pyburn, L., & Ames, A. J. (2016). *Madison Collaborative Annual Assessment Report #3*.

Solomonson, A. L. & Lance, C. E. (1997). Examination of the relationship between true halo and halo error in performance ratings. *Journal of Applied Psychology, 82,* 665-674.

Stecher, B. (2014). Looking back: Performance assessment in an era of standards-based educational accountability. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 17-52). San Francisco, CA: Jossey-Bass.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4).

Stiggins, R. J. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33-42.

Stone, C. A. & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS.* Cary, NC: SAS Institute, Inc.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9,* 239-261.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4,* 25-29.

Tierney, R. & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation, 9* (2).

Topol, B., Olson, J., Roeber, E. (2010). *The cost of new higher quality assessments: A comprehensive analysis of the potential costs for future state assessments.* Stanford, CA: Stanford Center for Opportunity Policy in Education.

United States Department of Education (2006). *A test of leadership: Charting the future of U.S. higher education.* Washington, DC: U.S. Department of Education.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practices, 15*(1), 22-29.

Wang, W-C. & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296-318.

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

Welch, C. (2006). Item and prompt development in performance testing. In Downing, S. M. & Haladyna, T. M. (Eds.). *Handbook of test development* (pp. 303-328). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Wendler, C. L. W. & Walker, M. E. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. M. Downing & T. M. Haladyna (Eds.). *Handbook of test development* (pp. 287-302). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Wiggins, G. (1991). A response to Cizek. *The Phi Delta Kappan, 72*, 700-703.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *The Phi Delta Kappan, 75*, 200-214.

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

Wiley, D. E. & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In M. B. Kane & R. Mitchell (Eds.) *Implementing performance assessment: Promises, problems, and challenges* (pp. 61-90). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*, 35-51.

Wolfe, E. W. & Chiu, C. W. T. (1997). Detecting rater effects with a multi-faceted rating scale model. Paper presented at The Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, March 1997.

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

Wolfe, E. W., Myford, C. M., Engelhard, G., Jr. & Manolo, J. R. (2007). Monitoring reader performance and DRIFT in the AP ® English Literature and Composition Examination using benchmark essays (Research Report 2007-2). New York, NY: The College Board.

Wright, B. D. & Masters, G. (1982). *Rating Scale Analysis.* Chicago, IL: Chicago Mesa Press.

Wu, S. M. & Tan, S. (2016). Managing rater effects through the use of FACETS analysis: The case of a university placement test. *Higher Education Research & Development, 35,* 380-394.

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31,* 501-527.

Table 1

*Rater $MS_U$ and $MS_W$ Estimates*

| Rater | $MS_U$ | $MS_W$ |
|---|---|---|
| 1 | 0.81 | 0.89 |
| 2 | 0.85 | 0.94 |
| 3 | 1.64 | 1.76 |
| 4 | 1.32 | 1.25 |
| 5 | 1.01 | 1.27 |
| 6 | 1.27 | 1.51 |
| 7 | 0.67 | 0.67 |
| 8 | 0.70 | 0.71 |
| 9 | 1.26 | 1.34 |
| 10 | 0.89 | 1.00 |
| 11 | 1.04 | 1.04 |
| 12 | 0.91 | 1.06 |
| 13 | 0.95 | 0.96 |
| 14 | 1.17 | 1.14 |
| 15 | 0.98 | 1.06 |
| 16 | 0.85 | 0.82 |
| 17 | 0.91 | 0.93 |
| 18 | 0.64 | 0.71 |

*Note.* $MS_U$ = unweighted mean square; $MS_W$ = weighted mean square.

Table 2

*Rater Leniency/Severity Estimates and Observed Score Descriptive Information*

| Rater | Total Ratings | Observed Score Sum | Observed Average Score | Leniency/Severity | *SE* |
|---|---|---|---|---|---|
| 3 | 260 | 543.00 | 2.09 | -1.52 | 0.05 |
| 7 | 290 | 456.50 | 1.58 | -1.06 | 0.05 |
| 8 | 285 | 368.50 | 1.30 | -0.62 | 0.05 |
| 2 | 285 | 328.00 | 1.15 | -0.28 | 0.05 |
| 15 | 295 | 359.50 | 1.22 | -0.20 | 0.05 |
| 16 | 295 | 336.00 | 1.14 | -0.07 | 0.05 |
| 11 | 260 | 313.00 | 1.21 | -0.03 | 0.06 |
| 18 | 265 | 305.50 | 1.16 | 0.03 | 0.06 |
| 1 | 285 | 273.50 | 0.96 | 0.07 | 0.06 |
| 10 | 295 | 275.50 | 0.94 | 0.10 | 0.06 |
| 9 | 295 | 273.00 | 0.93 | 0.12 | 0.06 |
| 17 | 265 | 290.50 | 1.10 | 0.13 | 0.06 |
| 12 | 255 | 268.50 | 1.06 | 0.25 | 0.06 |
| 13 | 275 | 334.50 | 1.22 | 0.37 | 0.05 |
| 6 | 280 | 268.50 | 0.96 | 0.54 | 0.06 |
| 4 | 275 | 210.50 | 0.77 | 0.63 | 0.06 |
| 14 | 270 | 267.00 | 0.99 | 0.73 | 0.06 |
| 5 | 280 | 231.50 | 0.83 | 0.80 | 0.06 |

*Note.* Observed score sum and observed average score are on the original 0 – 4 scale of the ER-WR rubric. Leniency/severity estimates are in logits. *SE* = standard error.

Table 3

*Rater by Element Interaction Results*

| | Element A | | | Element B | | | Element C | | | Element D | | | Element E | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Rater | Bias | *SE* | *t* | Bias | *SE* | *t* | Bias | *SE* | *t* | Bias | *SE* | *t* | Bias | *SE* | *t* |
| 1 | **0.30** | **0.11** | **2.76** | 0.07 | 0.12 | 0.60 | 0.01 | 0.14 | 0.05 | -0.30 | 0.15 | -1.92 | **-0.41** | **0.16** | **-2.55** |
| 2 | **-0.72** | **0.11** | **-6.73** | **0.34** | **0.11** | **3.04** | **0.32** | **0.12** | **2.66** | 0.17 | 0.13 | 1.37 | 0.11 | 0.13 | 0.81 |
| 3 | **-0.29** | **0.13** | **-2.24** | **-0.92** | **0.12** | **-7.80** | **-0.29** | **0.12** | **-2.46** | **0.53** | **0.12** | **4.56** | **0.91** | **0.12** | **7.59** |
| 4 | **-0.35** | **0.12** | **-2.96** | **0.44** | **0.13** | **3.37** | **0.33** | **0.15** | **2.23** | -0.35 | 0.18 | -1.99 | 0.00 | 0.16 | 0.03 |
| 5 | **0.40** | **0.11** | **3.78** | **-0.44** | **0.15** | **-2.98** | 0.19 | 0.14 | 1.31 | -0.23 | 0.16 | -1.40 | -0.32 | 0.17 | -1.92 |
| 6 | **0.77** | **0.11** | **6.91** | 0.01 | 0.13 | 0.07 | **-0.55** | **0.16** | **-3.43** | **-0.65** | **0.17** | **-3.84** | **-0.35** | **0.16** | **-2.22** |
| 7 | -0.20 | 0.11 | -1.83 | -0.03 | 0.11 | -0.31 | 0.03 | 0.11 | 0.27 | 0.16 | 0.11 | 1.44 | 0.06 | 0.11 | 0.53 |
| 8 | 0.06 | 0.11 | 0.52 | -0.04 | 0.11 | -0.34 | 0.16 | 0.12 | 1.38 | -0.02 | 0.13 | -0.18 | -0.19 | 0.13 | -1.45 |
| 9 | **0.57** | **0.11** | **5.19** | -0.21 | 0.14 | -1.53 | -0.11 | 0.15 | -0.73 | -0.25 | 0.15 | -1.59 | **-0.51** | **0.16** | **-3.11** |
| 10 | **0.22** | **0.11** | **2.07** | -0.22 | 0.14 | -1.66 | -0.06 | 0.15 | -0.41 | -0.06 | 0.15 | -0.39 | -0.02 | 0.15 | -0.15 |
| 11 | **-0.41** | **0.11** | **-3.79** | 0.21 | 0.12 | 1.76 | 0.09 | 0.13 | 0.66 | **0.37** | **0.13** | **2.92** | -0.11 | 0.14 | -0.78 |
| 12 | -0.17 | 0.11 | -1.55 | -0.02 | 0.13 | -0.17 | 0.02 | 0.14 | 0.15 | 0.15 | 0.14 | 1.06 | 0.14 | 0.14 | 1.01 |
| 13 | -0.01 | 0.11 | -0.06 | 0.15 | 0.11 | 1.32 | -0.03 | 0.13 | -0.22 | -0.15 | 0.13 | -1.11 | -0.01 | 0.13 | -0.11 |
| 14 | -0.11 | 0.11 | -0.99 | 0.17 | 0.12 | 1.40 | -0.18 | 0.15 | -1.19 | 0.04 | 0.14 | 0.26 | 0.09 | 0.14 | 0.65 |
| 15 | **0.39** | **0.11** | **3.69** | -0.19 | 0.12 | -1.67 | **-0.58** | **0.14** | **-4.04** | -0.08 | 0.13 | -0.65 | 0.24 | 0.12 | 2.00 |
| 16 | -0.04 | 0.10 | -0.41 | **0.29** | **0.11** | **2.75** | 0.25 | 0.12 | 2.03 | -0.15 | 0.14 | -1.09 | **-0.54** | **0.15** | **-3.62** |
| 17 | **-0.49** | **0.11** | **-4.65** | 0.13 | 0.12 | 1.09 | 0.21 | 0.13 | 1.60 | 0.15 | 0.13 | 1.15 | **0.29** | **0.13** | **2.22** |
| 18 | -0.08 | 0.10 | -0.79 | 0.15 | 0.11 | 1.36 | 0.10 | 0.13 | 0.75 | -0.05 | 0.14 | -0.36 | -0.13 | 0.14 | -0.95 |

*Note*. Bolded values indicate *t*-values statistically significant at $p < .05$. *SE* = standard error.

Table 4

*Frequencies of scores provided by raters in each score category*

| | Score | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Rater | 0 | 0.5 | 1 | 1.5 | 2 | 2.5 | 3 | 3.5 | 4 | Total |
| 1 | 77 | 81 | 47 | 20 | 21 | 13 | 22 | 1 | 3 | 285 |
| 2 | 53 | 56 | 68 | 39 | 36 | 10 | 12 | 7 | 4 | 285 |
| 3 | 29 | 19 | 33 | 27 | 35 | 24 | 31 | 22 | 40 | 260 |
| 4 | 93 | 86 | 37 | 20 | 18 | 7 | 8 | 2 | 4 | 275 |
| 5 | 102 | 72 | 43 | 12 | 17 | 11 | 18 | 2 | 3 | 280 |
| 6 | 117 | 53 | 32 | 14 | 12 | 5 | 30 | 11 | 6 | 280 |
| 7 | 21 | 31 | 86 | 41 | 32 | 32 | 35 | 7 | 5 | 290 |
| 8 | 52 | 50 | 70 | 23 | 34 | 16 | 25 | 8 | 7 | 285 |
| 9 | 119 | 38 | 60 | 14 | 21 | 9 | 25 | 5 | 4 | 295 |
| 10 | 125 | 35 | 57 | 9 | 28 | 9 | 18 | 2 | 12 | 295 |
| 11 | 31 | 67 | 66 | 22 | 34 | 21 | 14 | 4 | 1 | 260 |
| 12 | 57 | 65 | 57 | 17 | 18 | 21 | 10 | 10 | 0 | 255 |
| 13 | 58 | 60 | 45 | 34 | 21 | 22 | 25 | 7 | 3 | 275 |
| 14 | 72 | 41 | 82 | 16 | 34 | 9 | 13 | 2 | 1 | 270 |
| 15 | 48 | 80 | 81 | 12 | 17 | 9 | 23 | 10 | 15 | 295 |
| 16 | 39 | 78 | 82 | 25 | 30 | 20 | 15 | 3 | 3 | 295 |
| 17 | 25 | 68 | 84 | 39 | 32 | 7 | 6 | 3 | 1 | 265 |
| 18 | 26 | 73 | 88 | 28 | 14 | 13 | 9 | 9 | 5 | 265 |
| Total | 1144 | 1053 | 1118 | 412 | 454 | 258 | 339 | 115 | 117 | 5010 |
| Percent | 22.83 | 21.02 | 22.32 | 8.22 | 9.06 | 5.15 | 6.77 | 2.30 | 2.34 | |

*Note.* Shading represents rater pairs who rated the same essays (i.e. raters 1 and 2 were a pair, raters 3 and 4 were a pair, etc.). The total number of ratings varies by raters based on if their scores were removed due to "unrateable" essays or if their scores were removed due to failure to follow the special notes on the ER-WR rubric.

| | Student Scores | | | | |
|---|---|---|---|---|---|
| | Element A | Element B | Element C | Element D | Element E |
| Case 1: no special notes in effect | 4 | 4 | 3 | 3 | 4 |
| Case 2: special note for Element D in effect | 3 | 3 | 3 | 1 | 1 |
| Case 3: special notes for Elements D and E in effect | 3 | 4 | 1 | 1 | 0 |

| CASE | ELEMENT | SCORE |
|---|---|---|
| 1 | A | 4 |
| 1 | B | 4 |
| 1 | C | 3 |
| 1 | Dl | m |
| 1 | Du | 3 |
| 1 | El | m |
| 1 | Eu | 4 |

| CASE | ELEMENT | SCORE |
|---|---|---|
| 2 | A | 3 |
| 2 | B | 3 |
| 2 | C | 3 |
| 2 | Dl | m |
| 2 | Du | 1 |
| 2 | El | 1 |
| 2 | Eu | m |

| CASE | ELEMENT | SCORE |
|---|---|---|
| 3 | A | 3 |
| 3 | B | 4 |
| 3 | C | 1 |
| 3 | Dl | 1 |
| 3 | Du | m |
| 3 | El | 0 |
| 3 | Eu | m |

*Figure 1*. Example data structure with upper and lower elements. “m” indicates missing data.

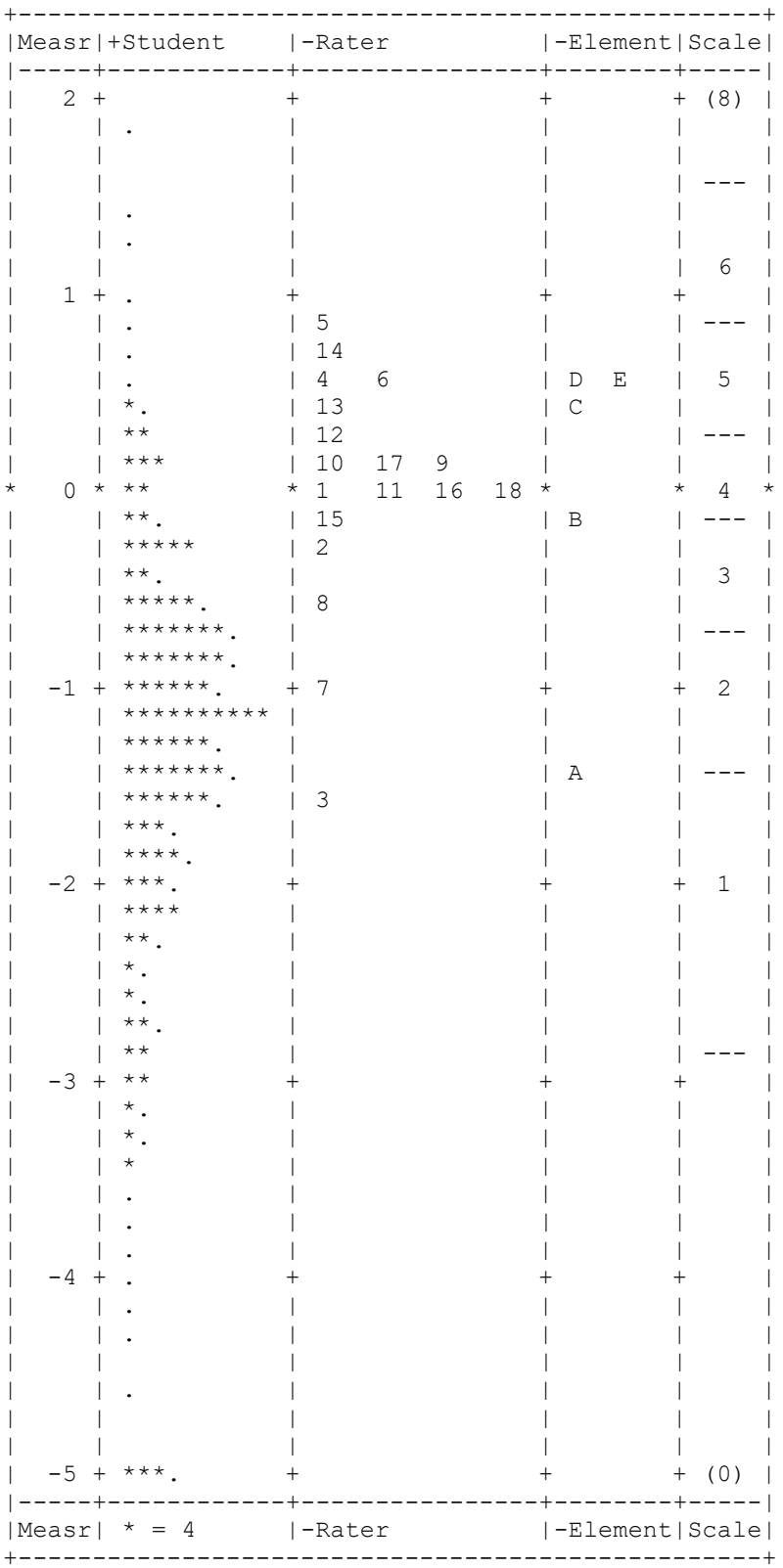


*Figure 2.* Wright Map generated in FACETS (Linacre, 2017b) output. The rater and element facets were centered at 0.00; the student facet was free to vary. The student facet was oriented positively, such that higher logits represent greater ability than lower logits.

The rater and element facets were oriented negatively, such that higher logits represent more severe raters and more difficult elements, respectively, compared to lower logits.

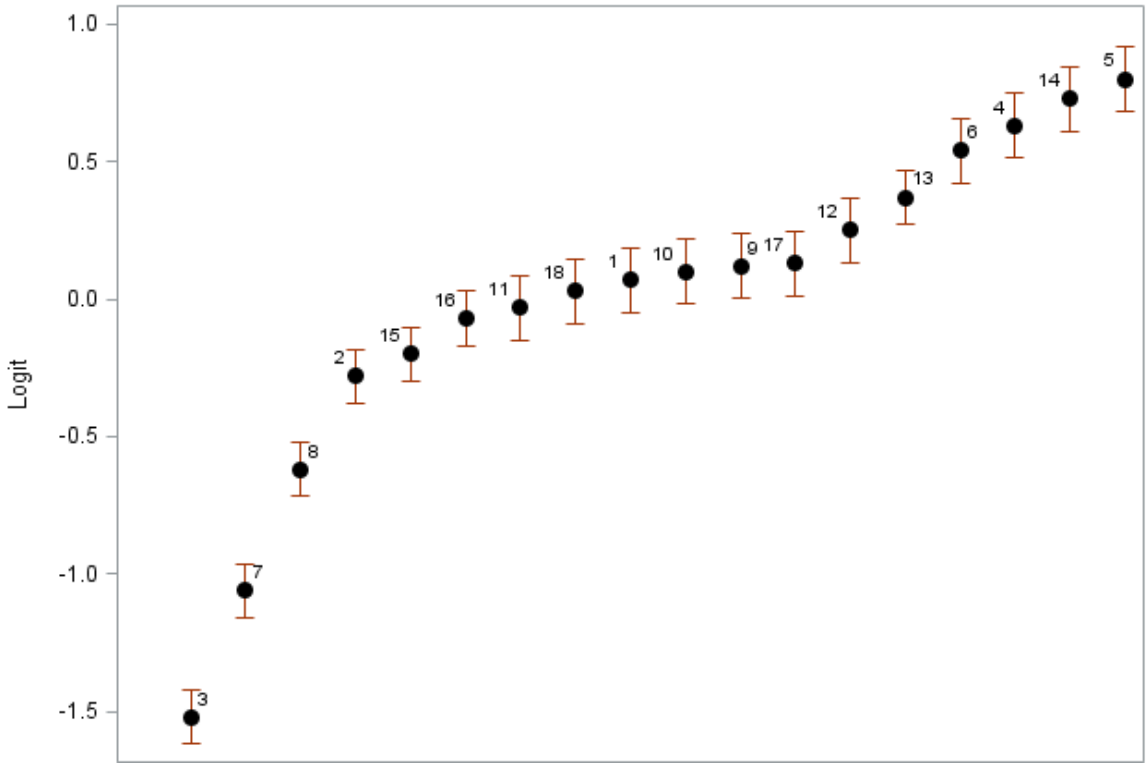


*Figure 3.* Confidence intervals for individual raters' leniency/severity logits. Numbers indicate raters.

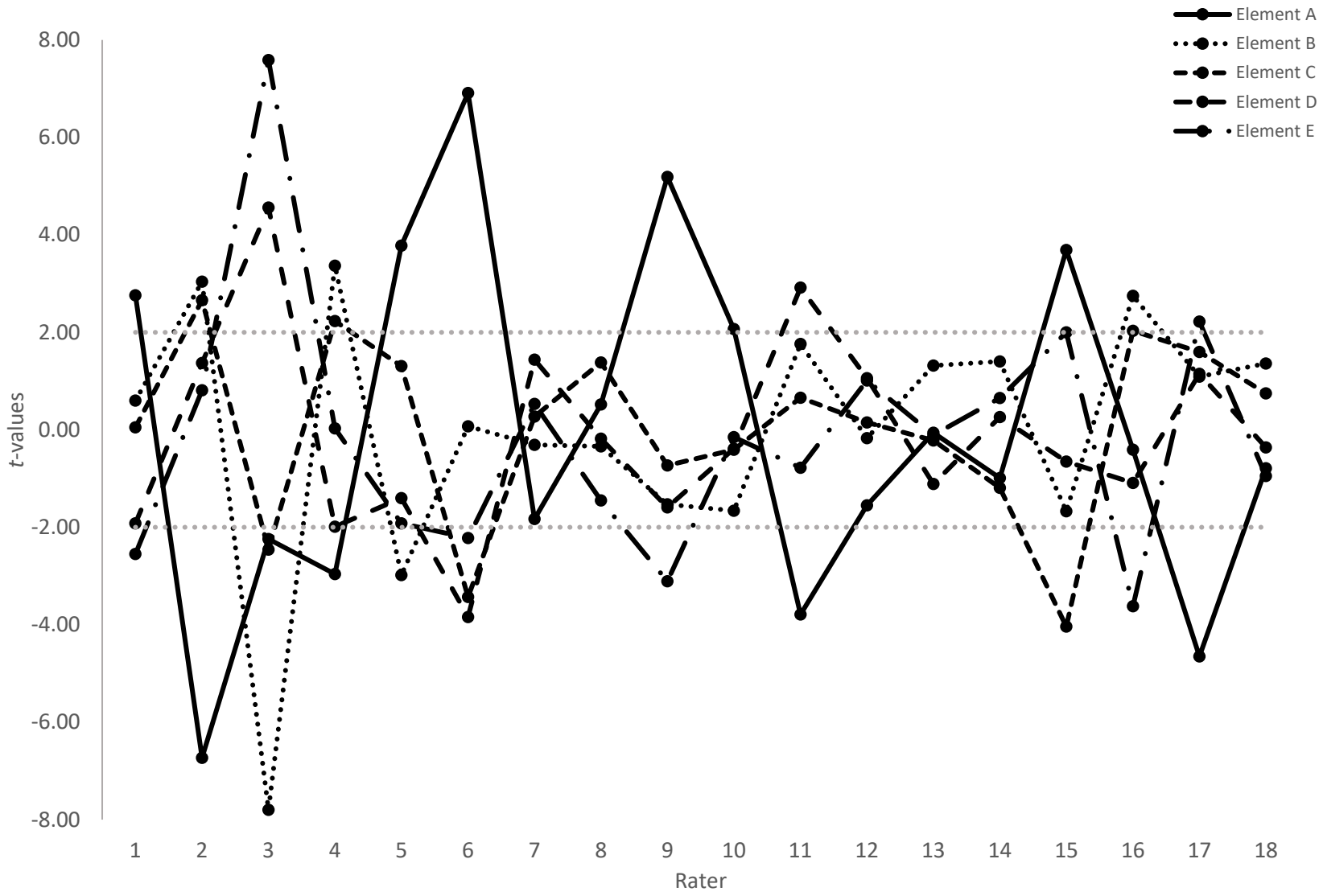


*Figure 4.* Rater by element interaction bias diagram. Estimates above 2 and below -2 indicate statistically significant t-values.

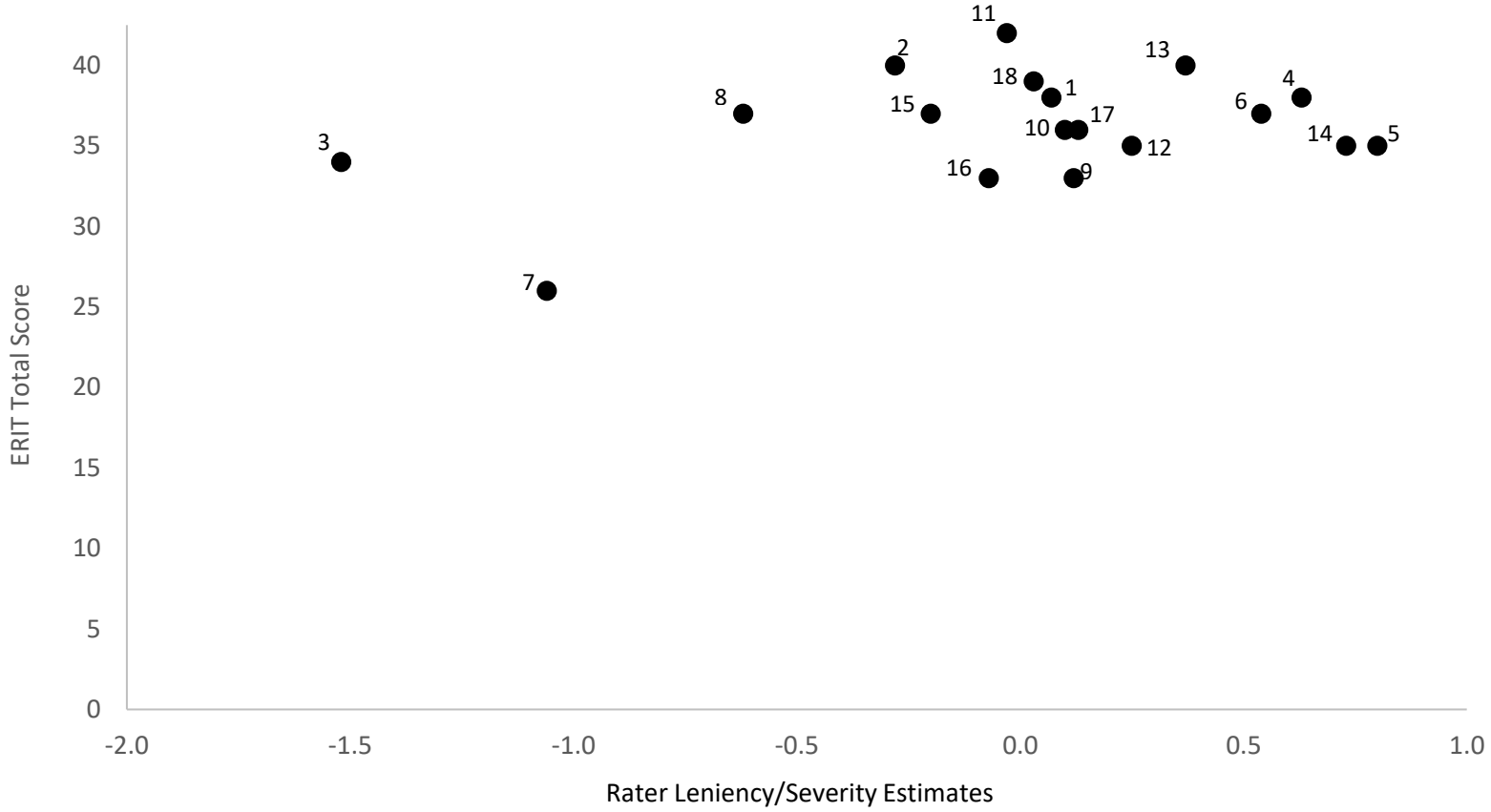


*Figure 5.* Correlation between raters' leniency/severity logits and their 8KQ knowledge as measured by the Ethical Reasoning Identification Test. Numbers indicate raters.

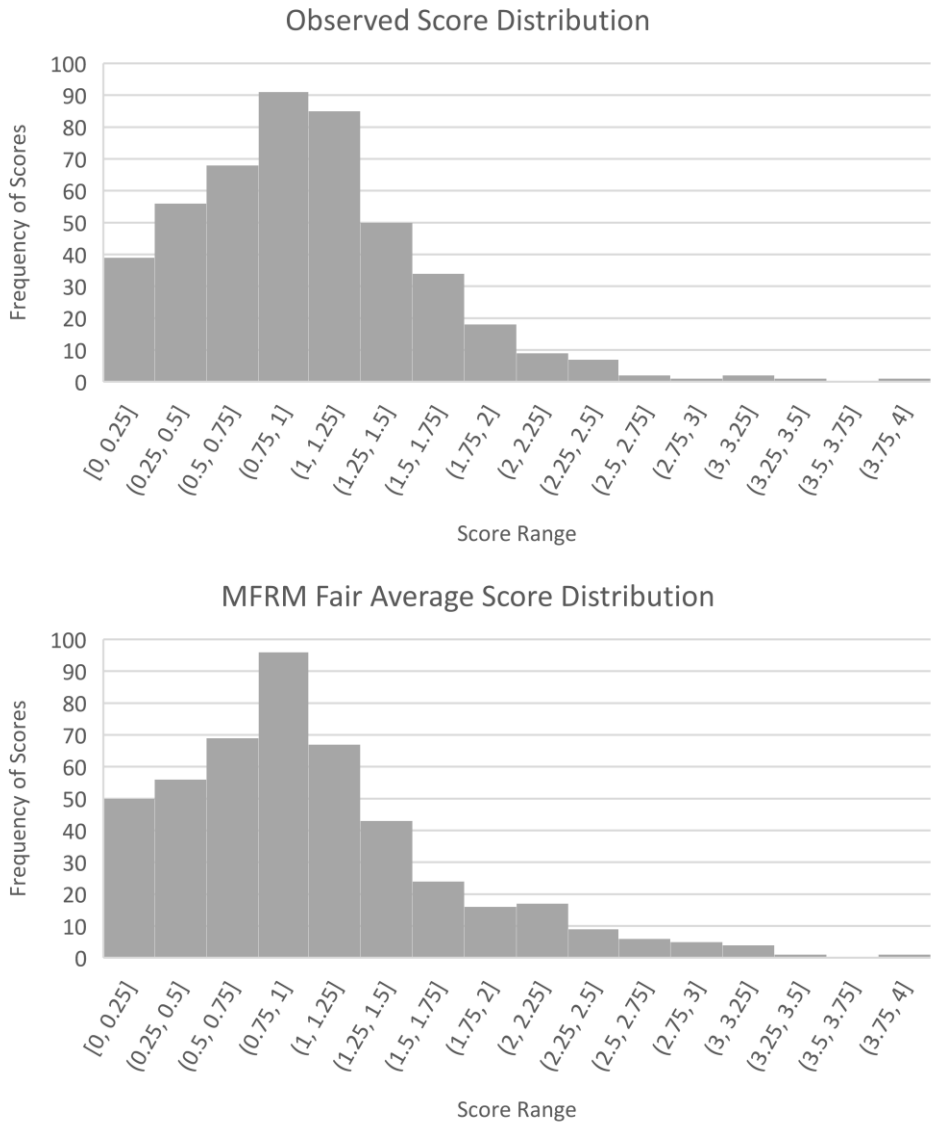


*Figure 6.* Observed score and MFRM fair average score histograms

# Appendix A

## ER-WR Prompt and Administration Instructions

***[Note to proctors:*** *non-lead proctor should pass out the Ethical Reasoning-WRA handout to students as the lead proctor reads the script (as well as scrap paper, if it has not already been passed out). Also, please note that this test has two parts, an essay and an additional 5 questions. Students will complete the essay together, and then you will instruct them to complete the additional 5 items. We encourage you to show students how to navigate to the 5 additional items using the projector.****]***

This test is an ethical reasoning assessment.

Often in life, we encounter situations that are 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 JMU are interested in the ethical reasoning thought process in which students engage when confronted with such situations.

For this assessment, please explain a complicated situation with which you are very familiar, the ethical thought process you used to address the situation, and the decision that was made.

You will have 55 minutes to compose this essay. Your document should be no fewer than 250 words. For your convenience, you are given a piece of paper that repeats the instructions for this task in more detail. You may refer to this piece of paper throughout this assessment. Additionally, you have been provided with scrap paper. You can use the scrap paper to outline your ideas, brainstorm, or apply any other technique to facilitate your writing.

Please feel free to express whatever opinions you might hold. Your essay will NOT be evaluated on what decision was chosen, but rather the clarity and complexity of the thought process underlying that decision.

You will be told when there are 10 and 5 minutes remaining.

On the assessment webpage, please click on the link for the ER-WRA assessment test.

Please fill in your JACard number at the top. Then select the ER-WRA test from the dropdown menu. Then write your first name, last name, and finally your JACard number again. Do not insert self-identifying information anywhere else on the screen.

**INSTRUCTIONS CONTINUE ON NEXT PAGE ⊳⊳⊳⊳⊳⊳**

Please save the document frequently as you type. In the event the program is accidentally closed, open up the assessment again, type in your JACard number at the top, and click the "Retrieve" button.

When you have finished, please stay on the essay page and sit quietly until the testing time is over. Again, please do NOT close out of this window.

Raise your hand if you have any trouble accessing the test.

You may begin.

***Remind students when they have 10 and 5 minutes remaining for the essay portion of this test. Once 55 minutes have elapsed, please instruct students to continue on to the 5 additional test items by reading the following aloud:***

Thank you for working on this important assessment. Before moving on to the next assessment, please be sure to save your document once again by clicking the "Save" button.

Once you have saved your document, please click on the link at the bottom of the essay page that says “Dosage” in order to complete a few more questions. You will have 5 minutes to complete these additional questions.

You may begin.

***[See below for a screenshot of where to find this link. If a student accidentally closes the essay page window they can still take the additional items. Simply have them click on the ER-WRA link from the main Assessment Day website to re-open the page. It may be helpful for you to show students where the link to the extra items is using the classroom projector. Instructions for retrieving essays are also included below.]***

***[Note to proctors:*** *non-lead proctor should collect the ER-WRA handouts from the students after they have completed the assessment. Please place all ER-WRA handouts in the designated envelope within your bin.****]***

# Appendix B

## ER-WR Rubric

*James Madison University's Ethical Reasoning Rubric*

| Insufficient<br>0 | Marginal<br>1 | Good<br>2 | Excellent<br>3 | Extraordinary<br>4 | Score |
|---|---|---|---|---|---|
| **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…<br>• Context treated with nuance<br>• 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…<br>Nuanced treatment of key questions, for example:<br>• elucidates subtle distinctions<br>• uses analogies or metaphors<br>• 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….<br>Logically terminates in decision that will be reached. | Meets criteria for *Excellent* AND…<br>Products of analysis weighed to make judgment compelling. | |

# Appendix C

## Rating Timeline

| **Day 1** | |
|---|---|
| 8:30am | Raters report for breakfast |
| 9:00am – 11:30am | 8KQ Workshop |
| 11:30am – 12:00pm | Raters complete the ERIT |
| 12:00pm – 12:30pm | Lunch |
| 12:30pm – 2:30pm | ER-WR rubric training |
| 2:30pm – 4:30pm | Raters rate |

| **Day 2** | |
|---|---|
| 8:30am | Raters report for breakfast |
| 9:00am – 10am | ER-WR rubric refresher training |
| 10:00am – 12:00pm | Raters rate |
| 12:00pm – 12:30pm | Lunch |
| 12:30pm – 4:30pm | Raters rate |

# Appendix D

## 8KQ synonyms

**Assessment Rubric Examples (8KQ synonyms)**

If the response uses the key term **EMPATHY** or related terms **care, love, feelings** or provides a rationale involving personal sentiment for other persons or sentient beings, award X points.

If the response uses the key term **FAIRNESS** or related terms **justice, equality, balancing all legitimate interests**, or provides a rationale involving objective consideration of all ethical considerations without discrimination or prejudice, award X points.

If the response uses the key term **CHARACTER** or related terms **integrity, virtue, actualized self, ideal self, self-respect, the person that I am or would be, being ashamed or unable to live with one's self,** or provides a rationale involving reference to themselves, what they believe they are and/or the kind of person they desire to be, award X points.

If the response uses the key term **RIGHTS** or related terms **entitlements, dignity, respect-worthy, Bill of Rights, Universal Declaration of Human Rights,** or provides a rationale involving viewing others as deserving to be treated with respect or reverence, award X points.

If the response uses the key term **LIBERTY** or related terms **freedom, autonomy, consent,** or provides a rationale involving consideration of the choices, judgments or decisions of other persons, award X points.

If the response uses the key term **RESPONSIBILITY** or related terms **duty, obligation, debt, reciprocity,** or provides a rationale that refers to what others are owed because they are human beings, or promises, or special relationships with the person deciding what to do, award X points.

If the response uses the key term **OUTCOMES** or related terms **results, effects, consequences, utility, preferences, happiness, greatest good for the greatest number, Karma,** or provides a rationale involving reference to some calculation, prediction or anticipation about the value of what comes about as a result of one's choice or action, award X points.

If the response uses the key term **AUTHORITY** or related terms **command, orders, legally required, God's commands or requirements,** or provides a rationale involving reference to having to do what they are going to do because some other person or institution requires it of them, award X points.

W. J. Hawk 5 December 2012

# Appendix E

## Equations

| Equation | Notation | Equation Number |
|---|---|---|
| MFRM Rating Scale Model | $ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_k$ | 1 |
| MFRM Partial Credit Model | $ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_{ijk}$ | 2 |
| MFRM Hybrid Model | $ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_{ik}$ | 3 |
| Fixed-effect Chi-square | $x^2 = \sum(w_o * D_o^2) - \frac{(\sum w_o * D_o)^2}{\sum w_o}$ | 4 |
| True Standard Deviation | $SD_t^2 = SD_o^2 - MSE$ | 5 |
| Separation Ratio | $G_o = \sqrt{\frac{SD_t^2}{MSE}}$ | 6 |
| Separation Index | $H_o = \frac{4\sqrt{\frac{SD_t^2}{MSE}} + 1}{3}$ | 7 |
| Reliability of Separation | $R_o = \frac{\frac{SD_t^2}{MSE}}{1 + \frac{SD_t^2}{MSE}}$ | 8 |
| Standardized Residual | $Z_{nij} = \frac{x_{nij} - e_{nij}}{\sqrt{w_{nij}}}$ | 9 |
| Expected Rating | $e_{nij} = \sum_{k=0}^{m} kp_{nijk}$ | 10 |
| Model Variance | $w_{nij} = \sum_{k=0}^{m}(k - e_{nij})^2 p_{nijk}$ | 11 |
| Unweighted Mean Square/Outfit | $MS_{U_j} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} Z_{nij}^2}{NI}$ | 12 |
| Weighted Mean Square/Infit | $MS_{W_j} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} z_{nij}^2 w_{nij}}{\sum_{n=1}^{N}\sum_{i=1}^{I} w_{nij}}$ | 13 |

| | | |
|---|---|---|
| 95% Confidence Interval | Rater logit $\pm$ 1.96($SE_{rater}$) | 14 |
| MFRM Hybrid Model with Rater by Element Interaction | $ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \varphi_{ij} - \tau_{ik}$ | 15 |
| Rater by Element Bias Parameter | $t_{ij} = \frac{\hat{\varphi}_{ij}}{SE_{ij}}$ | 16 |

# Appendix F

Element location and threshold estimate comparisons for five- and seven-element models

Table F1

*Reliability Estimates for the Five- and Seven-Element Models*

| Model | Student Reliability of Separation | Rater Reliability of Separation |
|---|---|---|
| Five-element | 0.88 | 0.99 |
| Seven-element | 0.87 | 0.99 |

*Note.* Five-element model included Elements A, B, C, D, and E. Seven-element model included Elements A, B, C, $D_{lower}$, $D_{upper}$, $E_{lower}$, and $E_{upper}$.

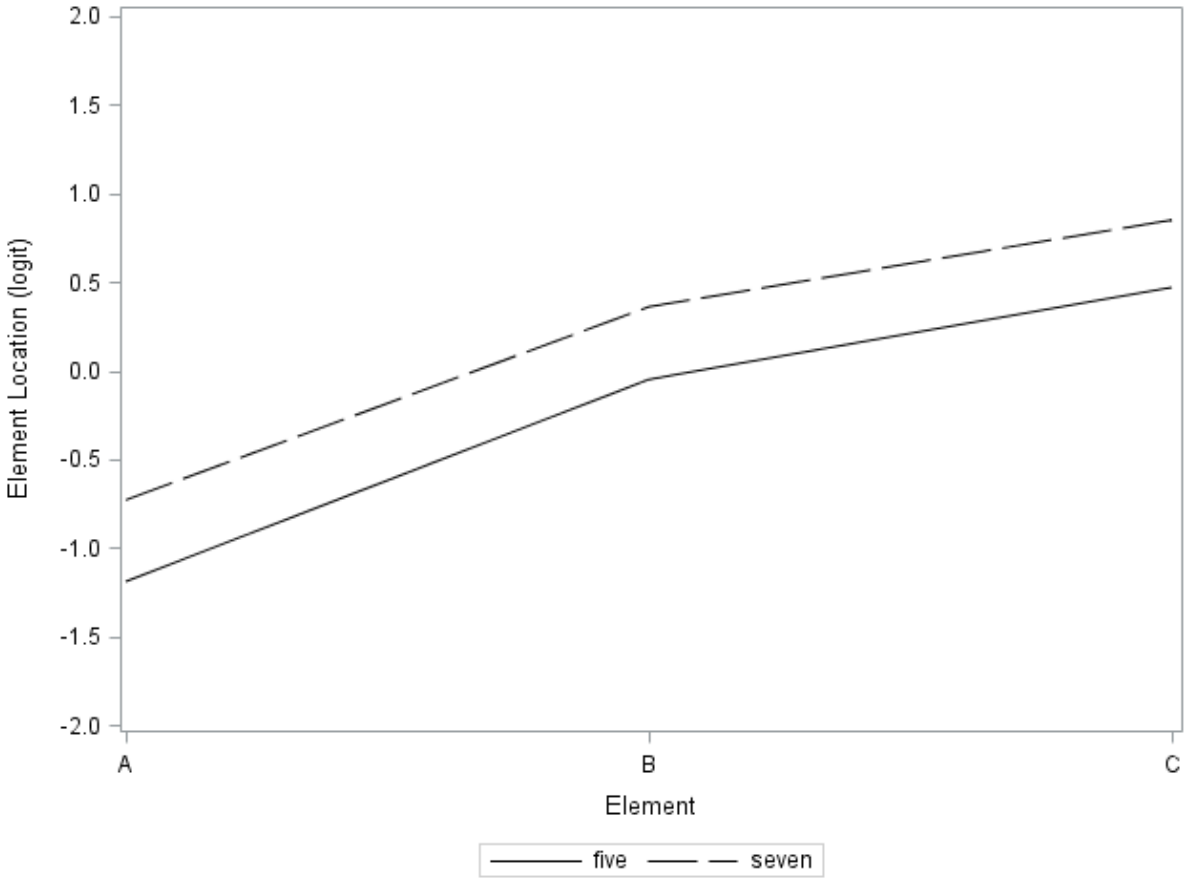


*Figure F1.* Item location estimates for the five- and seven-element models.

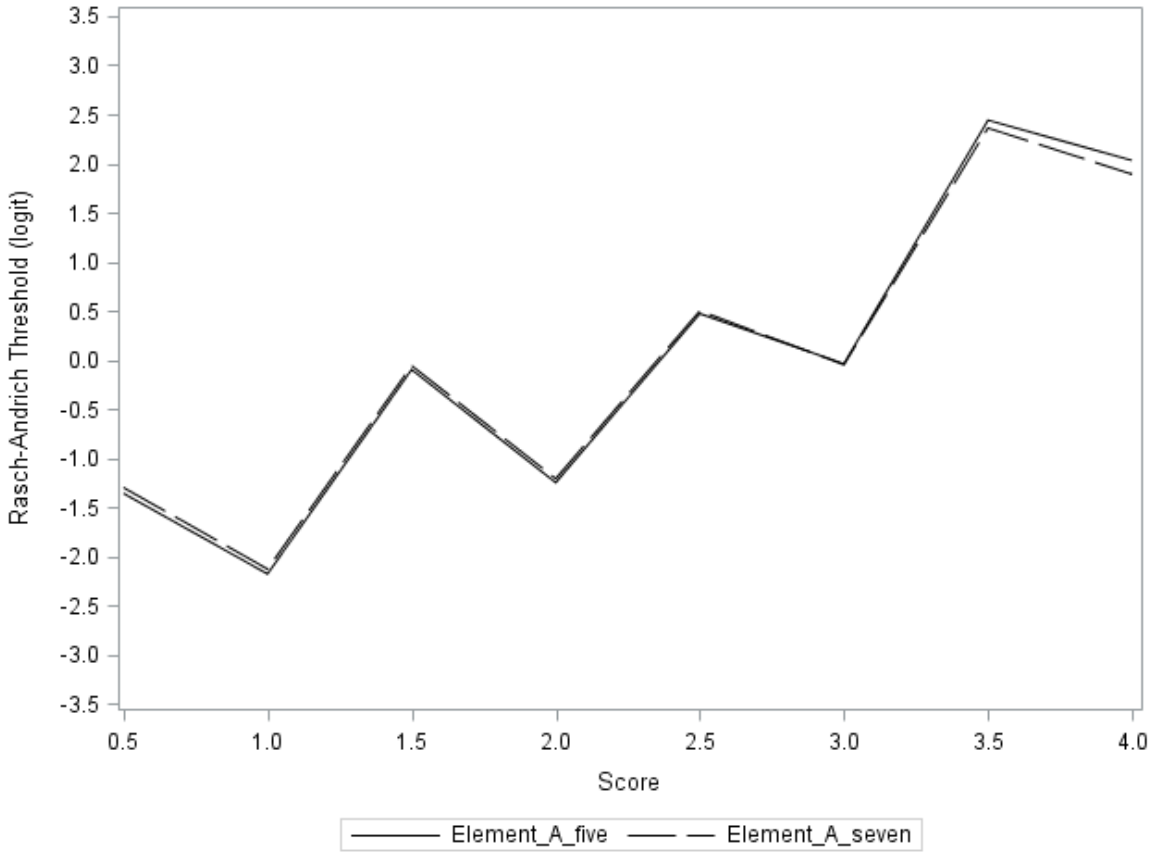


*Figure F2.* Rasch-Andrich threshold estimates for Element A for five- and seven-element models

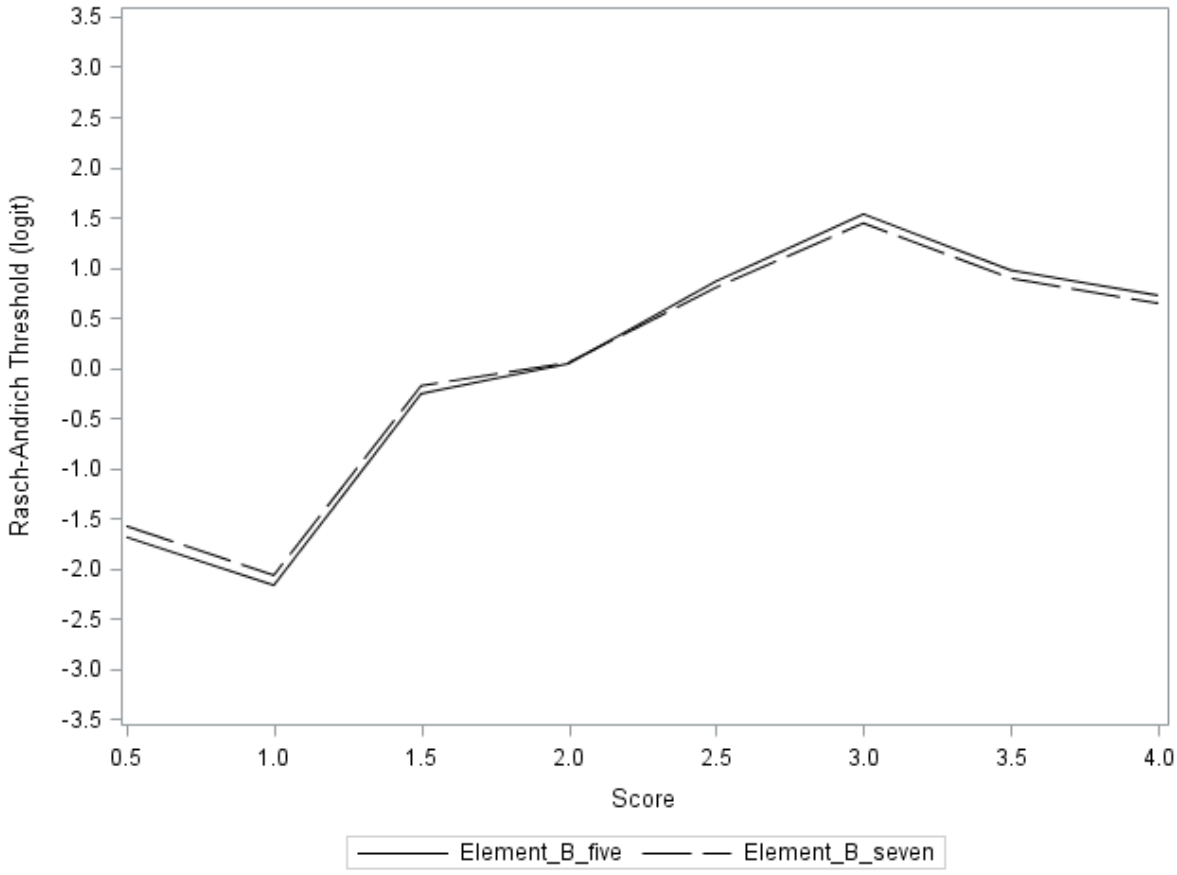


*Figure F3.* Rasch-Andrich threshold estimates for Element B for five- and seven-element models.

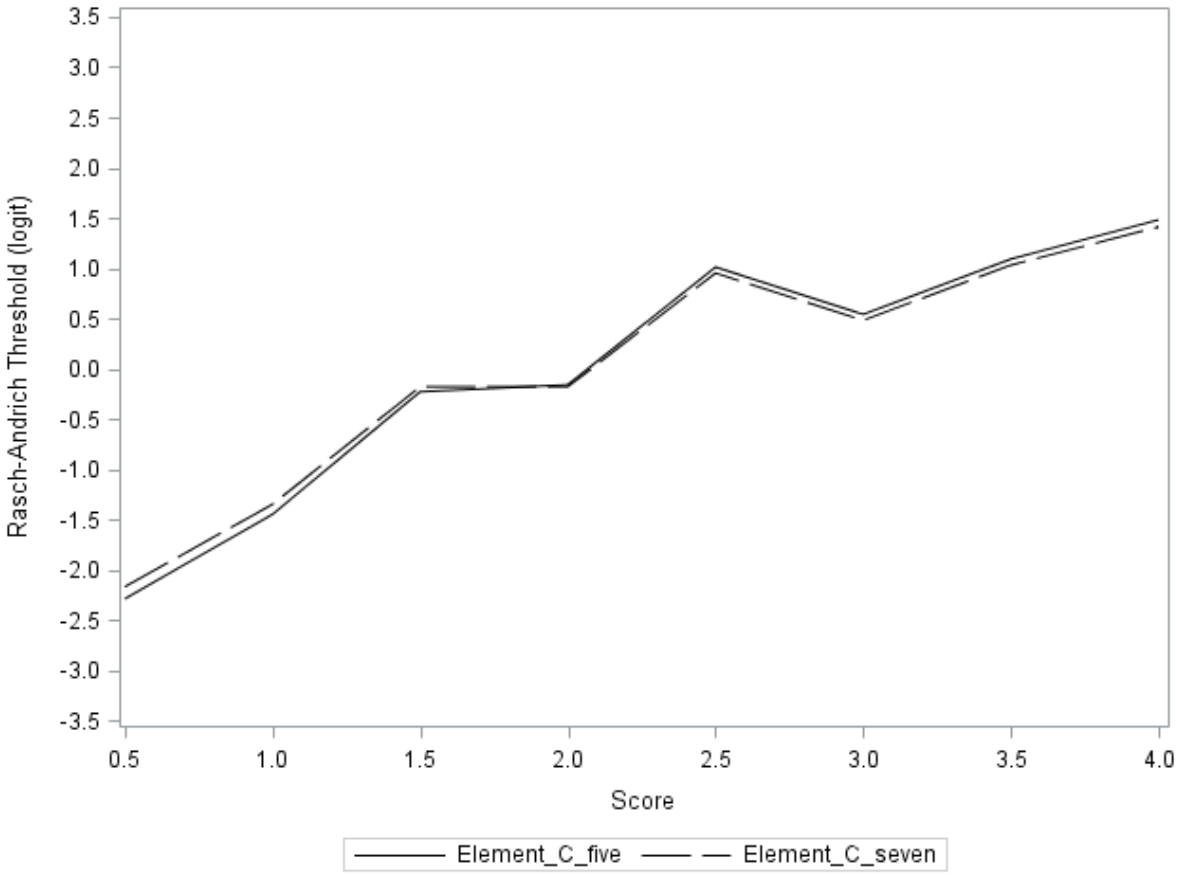


*Figure F4.* Rasch-Andrich threshold estimates for Element C for five- and seven-element models.