5-6-2021

Identifying rater effects for writing and critical thinking: Applying the Many-Facets Rasch Model to the VALUE Institute

Yelisey A. Shapovalov
James Madison University

Identifying Rater Effects for Writing and Critical Thinking: Applying the Many-Facets Rasch Model to the VALUE Institute

Yelisey A. Shapovalov

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

For the degree of

Master of Arts

Department of Graduate Psychology

May 2021

FACULTY COMMITTEE:

Committee Chair: John Hathcoat

Committee Members:

Christine DeMars

Jeanne Horst

Kathryne McConnell

Acknowledgements

A special thanks to all of my thesis committee members: John, who's guidance and wisdom throughout the processes was invaluable; Christine, for her diligent and keenly useful feedback; Jeanne, who never failed to be encouraging and insightful; and Kate, for extending a warm welcome and research opportunities from AAC&U, as well as providing me with a wealth of data.

I am grateful to my parents, for their support, although likely not quantifiable, was statistically significant. Thank you to my siblings, friends, and fans whose prayers were not in vain.

Table of Contents

List of Tables

List of Figures

Abstract

Performance assessments require examinees to carry out a process or produce a product and can be designed to have high fidelity to real-world application of higher-order skills. As such, performance assessments are highly valued in higher education settings. However, performance assessment is vulnerable to psychometric challenges that threaten the validity of scores due to the subjective nature of the scoring process. Specifically, raters must exercise judgement to provide scores to examinee work, which may be impacted by rater effects, or systematic differences in how raters evaluate performance assessment artifacts. Research has indicated that performance assessment may never be fully free from errors in rater judgement. Consequently, additional quality control measures are investigated in the hopes of reducing the impact of rater effects by selecting raters that have not exhibited rater effect in previous performance assessment assignments. The purpose of this project was to evaluate VALUE Institute artifact scores for diagnostic information of rater effects. The Many-Facets Rasch Measurement (MFRM) model was used to evaluate VALUE Institute scores for rater leniency/severity effects, halo effect, and restriction of range effect. Data for the 2018-2019 academic year was collected by the VALUE Institute of the Association of American Colleges and Universities (AAC&U) on two of their most popular VALUE (Valid Assessment of Learning in Undergraduate Education) Rubrics: Critical Thinking and Written Communication. A series of follow-up evaluation of MFRM indices were conducted to identify which raters were exhibiting rater effects to create a pool of preferable raters for selection who did not exhibit rater effects. Findings showed that only a few raters exhibited rater effects, building confidence in the validity of scores produced by the

VALUE Institute using the VALUE Rubrics. Moreover, MFRM methods were successful in flagging initial raters for rater effects. Mixed success was experienced with follow-up frequency procedures to confirm how raters assign scores, suggesting a limitation of relying solely on frequency counts to identify rater effects. Recommendations for future research are made and the subjectivity of judgement in MFRM interpretation and classification is discussed. Ultimately, preferable raters were identified by using MFRM diagnostic information flagging raters exhibiting rater effects.

Chapter 1: Introduction

The first modern Olympics took place in 1896 in Athens, and featured 280 participants from 13 nations, competing in 43 events to a crowd of 60,000 and King Georgios I — the king of Greece (History.com Editors, 2018). Since 1994, the Summer and Winter Olympic Games have been held separately and have alternated every two years. Over a quarter of the world population, 1.92 billion people, watched the broadcast coverage of Pyeongchang Olympic Winter Games of 2018 (Gough, 2020). In between Olympic seasons, the International Sports Federation were selecting candidates to nominate as judges for upcoming the Olympics (Holter, 2018). The nominees must then be accepted by the International Olympic Committee (IOC) to officially serve as judges. Selected judges were responsible for rating the performances of athletes in competitions like figures skating.

How did the International Sports Federation boards determine who to nominate? How did the International Skating Union decide who should judge figure skating for the Olympics? They needed to select judges that accurately interpret athletes' performances. In other words, the judges must be able to evaluate the quality of figure skating without bias, or systematic errors in judgement. Judges must be consistent in the quality of their ratings over long periods of time. For example, they should not grow more severe or lenient if they become fatigued. Additionally, their judgements should not be impacted by characteristics of the athletes that are irrelevant to their performance. Ultimately, valid determination of the best figure skater in the world comes down to the accuracy and fairness of the judgements made by raters; therefore, selecting *who* should be a judge is of paramount importance.

The American Association of Colleges & Universities (AAC&U) is also interested in the rater selection question. AAC&U is an organization that seeks to improve undergraduate education. One way in which AAC&U aims to accomplish this goal is by providing quality assessment of higher order skills like critical thinking and written communication (AAC&U, 2019). In 2009, AAC&U released 16 rubrics to guide assessment of student work as part of the VALUE project, the Valid Assessment of Learning in Undergraduate Education. These VALUE rubrics were designed to be applied to authentic work samples embedded in undergraduate courses. Eventually, due to the popularity of VALUE rubrics, AAC&U launched the VALUE Institute where higher education institutions could send student work samples to be rated by VALUE-certified raters. However, AAC&U must decide who should rate student work samples for the VALUE Institute. Moreover, of the people hired as raters before, *who* should be called back to serve as raters again and by *what criteria* should rater selection decisions be made?

In the present study, I explore a rater selection method by evaluating the quality of raters' past judgements. First, I review the important role of assessment in higher education, particularly through the lenses of accountability and improvement. Next, I present the benefits and limitations of performance assessments. Specifically, performance assessments can tap into higher order skills, considered essential by many employers and postsecondary education programs. However, performance assessment scores are susceptible to errors in rater judgements due to the subjective nature of the scoring process. Then, I cover the quality control techniques typically employed to limit

the impact of rater effects. Finally, I describe the context of AAC&U and the need for an additional quality control method that can be used in the rater selection process.

Assessment in Higher Education

Assessment is an integral part of higher education. For the better part of a century, educational assessment has been focused on student learning outcomes, either of academic degree programs or institutional goals, often met in large part through a general education program. Student learning outcomes specify observable and measurable actions that students must be able to demonstrate. Calls for accountability, as heard in the Spellings Report released by the U.S. Department of Education's (2006) Commission on the Future of Higher Education, have set an even greater emphasis on assessment.

Assessment of student learning outcomes helps meet accreditation requirements, which carries out four important roles. Modern accreditation primarily serves as quality assurance, signaling to students and the public that an institution or program meets at least threshold standards (Eaton, 2009). This in turn builds confidence in higher education among the private sector. Logistically, accreditation facilitates the transfer of credits and is required for access to federal funds such as student aid and other federal programs. In addition to assisting accountability, assessment is also important for institutional and programmatic improvement.

Leaders of the field are putting more emphasis on using assessment results for actionable change leading to improvement, thus closing the assessment loop (Banta & Blaich, 2011). Assessment for improvement seeks to identify where a program is deficient and respond with formative change in order to bring up student performance on those learning outcomes (Ewell, 2009). Without assessment, educators cannot gauge the

efficacy of their programing and cannot take corrective action. However, the quality of information gained from assessment depends largely on instrumentation.

Assessing Higher Order Skills

Many higher education assessments employ selected-response formats, like multiple-choice, matching, and true-false. Assessment practitioners opt to use such selected-response assessments because they can cover a large breath of content and feature straightforward scoring procedures (Downing, 2006; Gronlund, 2003; Linn et al., 1991; Madaus & Kellaghan, 1993). However, selected-response assessments may not be an optimal means to gauge student learning on higher order knowledge, skills, and/or abilities. Learning outcomes of academic degree programs, and education institutions in general, often aim to develop students' higher order skills, such as critical thinking and written communication, which are typically better suited for performance assessment (Chickering, 1999; Lane & Stone, 2006; Wiggins, 1991). Performance assessments employ an open-response format requiring students to produce a product or engage in a process.

Consequently, higher education is experiencing a push encouraging the use of performance assessments to evaluate student learning objectives. Part of this push stems from the criticism of the Spellings Report (US Department of Education, 2006). Specifically, there is a concern that students do not fully develop the knowledge and skills required to be successful in the workforce upon graduation. Subsequent research supports this claim that graduates are lacking the cross-discipline, higher order knowledge and skills expected of new hires to perform and adapt to on-the-job demands (Arum & Roksa, 2011; Hart Research Associates, 2015). Moreover, assessment leaders

and faculty prefer information-rich, meaningful performance assessments, which are sometimes called authentic measures due to their tendency to reflect real-life tasks (Banta, Griffin, Flateby & Kahn, 2009). Through performance assessments, students are able to demonstrate complex skills. Consequently, educators can use performance assessment data to show quality of learning related to higher order outcomes; thereby satisfying external accountability requirements and meeting internal programmatic and institutional learning standards.

Rater Challenges in Performance Assessments

Despite the benefits of performance assessments and their popularity in modern higher education assessment, they are more susceptible to psychometric challenges than selected-response assessments. The burden of rectifying these issues hinders the widespread adoption and use of performance assessments. Two of the most limiting psychometric challenges relate to the reliability and validity of performance assessment scores. Reliability has to do with the reproducibility, or dependability, of assessment scores (Bandalos, 2018). High reliability means that the score a person obtained in a particular testing situation is consistent (or at least very similar) to the score they would obtain in another testing situation. Unreliability can stem from a lack of information because test scores are based on limited samples of behavior. Performance assessments are usually based on smaller samples of behavior than selected-response assessments (Traub & Rowley, 1991). This is because performance assessment tasks require considerably more time and resources to be completed and scored. Thus, performance assessments tend to produce scores that are be less reliable, or less consistent, than

selected-response assessments. Moreover, scores need to be reliable in order also be consider valid.

Validity refers to the "degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA & NCME, 2014, p. 11). A valid score interpretation represents the intended construct well, without interference of construct-irrelevant variance. Construct-irrelevant variance refers to the "degree to which test scores are affected by processes that are extraneous to the test's intended purpose" (AERA, APA & NCME, 2014, p. 12). For instance, test scores can be systematically impacted by processes that are not part of the construct. One of the more significant sources of construct irrelevant variance stems from the subjective nature of the scoring process in performance assessments. The products students create and processes they engage in for performance assessments typically need to be scored by human raters. These raters must exercise judgment to determine the extent to which students meet pre-specified scoring criteria, usually outlined in a rubric. On the other hand, selected-response assessments do not require human raters and are considered to have more objective scoring procedures. Typically, a correct response option is provided, and the scoring process consists of identifying whether students selected this correct response option. Thus, on selected-response assessments, humans act as scantrons where scores do not depend on who grades student responses.

Because performance assessment scores are rater-mediated, scores are potentially a product of rater idiosyncrasies in addition to, or instead of, student ability (Engelhard, 2002). Construct-irrelevant variance from raters threatens score validity of performance assessments. Consequently, performance assessment users must provide evidence that

scores are a function of student ability, not a function of raters (AERA, APA, & NCME, 2014). Ideally, raters would be interchangeable such that the score given to a student's product or process would be the same regardless of which rater mediated their score. Unfortunately, evidence suggests that rater effects impact performance assessment scores, resulting in weaker psychometric quality of scores and undermine score validity (Cizek, 1991a).

Scullen, Mount, and Goff (2000) defined rater effects as a "broad category of effects [resulting in] systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee" (p. 957). Three of the most prominent, well researched rater effects include leniency/severity, halo, and restriction of range (Myford & Wolfe, 2003). For instance, researchers have found that some raters tend to be more severe, giving lower scores across students, while others tend to be more lenient, inflating scores for all students. Such scores are not only representative of student ability but also of rater severity. This rater effect decreases the psychometric quality of scores by threatening the validity of score-based inferences. For example, if two students receive the same score but one was judged by a severe rater while the other by a lenient rater, then it would be inappropriate to infer that these two students possess the same ability.

Similar problems arise due to raters exhibiting halo or restriction of range tendencies. A rater who allows one characteristic of the product to impact judgements on separate, distinct dimensions of that product would be exhibiting a halo effect. Imagine for instance, a rater who reads an essay that answers a prompt correctly. However, the response is riddled with grammatical and spelling errors, which are part of a distinct

grading criterion. If the rater gives this essay a negative evaluation on the accuracy criteria due to the poor grammar and spelling, then the rater would be exhibiting a halo effect. Restriction of range effects are seen in raters who do not use the full range of the rating scale. For example, when judging presenters on a one to ten confidence scale, they may be hesitant to give scores lower than a seven. Such a rater would be exhibiting a restriction of range effect toward the upper end of the scale. Rubrics and rater training are the two most common methods employed to overcome the challenges posed by the subjective nature of performance assessments and the rater effects that may arise.

Rater Quality Controls for Valid Interpretations and Uses of Scores

Rubrics are scoring guides containing pre-specified criteria per score level. Rubrics help anchor the scoring process of performance assessments in some objectivity. Raters reference rubric criteria to identify the performance level exhibited in a product or process. However, raters must still exert judgement to match the student product or process to a rubric. Ideally, all raters would interpret the scoring criteria as intended by rubric developers and apply them consistently across ratees. Numerous rater training programs have been designed to improve quality of ratings and alignment in an effort to reduce rater errors — albeit with varying degrees of success (Bernardin & Pence, 1980; Hedge & Kavanagh, 1988; Landy & Farr, 1980; Latham, Wexley, & Pursell, 1975; McIntyre, Smith, & Hassett, 1984; Spool, 1978).

Both, rubrics and rater training, are developed and conducted prior to the scoring processes and both have widespread adoption where assessment performance is conducted. However, once the scoring process is complete, not much attention is given to rater data. The primary purpose for collecting the scores is to evaluate student work, but

these scores can also provide information about rater quality. Analysis of rater data can provide diagnostic information regarding which raters are exhibiting rater effects. Subsequently, informed decisions on which raters to invite back can be made in order to select raters that preserve the psychometric quality of scores.

AAC&U and the Need for Improved Rater Selection Methods

The American Association of Colleges & Universities (AAC&U), one of the most prominent advocates for the use of performances assessment in higher education, is keenly interested in how rater data can be used to identify which raters to select for subsequent scoring tasks. Having a methodology for identifying returning raters would be particularly useful for one of AAC&U's main projects. In 2009, AAC&U released the Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics to facilitate the use of performance assessments in higher education (AAC&U, 2019). Sixteen VALUE rubrics were developed to assess essential higher order skills, such as critical thinking and written communication. AAC&U designed the VALUE rubrics so that they could be adapted for use in various classroom setting to assess course-embedded student work. VALUE rubrics can also serve as large-scale assessment tools to summatively evaluate students' abilities to meet learning objectives related to higher order skills.

AAC&U's VALUE project has been successfully accepted by many educators and institutions. Within the first two years following their initial release, over 17,000 new individuals visited the website where the VALUE rubrics are freely available (AAC&U, 2019). In 2013, over 70 two- and four- year public institutions from 13 states submitted student work samples from their curricula to be rated on three VALUE rubrics by AAC&U trained scorers. The aggregated results of this collaborative initiative "provided

normed evidence of the quality landscape of student learning across institutions and states for external stakeholders, while also giving faculty helpful information for improving teaching and learning in courses and programs" (AAC&U, 2019). Due to growing interest in such assessment information, AAC&U worked to develop and launch the VALUE Institute in 2017.

The VALUE Institute is a resource "enabling any higher education institution, department, program, state, consortium or provider to utilize the VALUE rubrics approach to assessment by collecting and uploading samples of student work to a digital repository and have the work scored by certified VALUE Institute faculty and other educator scorers for external validation of institutional learning assessment" (AAC&U, n.d.). In order to provide accurate feedback, the VALUE Institute needs to employ certified raters who do not exhibit rater effects that can compromise score inferences. Given that VALUE Institute scores are used to make institution-level or program-level inferences regarding students' learning on complex abilities, it is warranted to further investigate VALUE Institute scores for rater effects and make decisions on which raters to select for future rating assignments based on this information.

Study Purpose & Research Questions

The purpose of the current study is to evaluate VALUE Institute scores for rater effects. Of specific interest is the evaluation of diagnostic rater leniency/severity, halo effect, and restriction of range. Moreover, the raters not exhibiting such rater effects can be recommended for future rating assignments.

Evaluating VALUE Institute scores for rater effects and identifying raters who do not exhibit rater effects will be useful for the VALUE Institute. If scores are not

influenced by rater effects, then this study would provide further validity evidence to support VALUE Institute score inferences on student abilities. If scores are influenced by rater effects, then this study provides further information regarding rater behaviors and identifies which raters are preferential for future rating assignments. This information is useful for the VALUE Institute, as it may have implication for the interpretations of VALUE Institute scores, as well as rater training and selection of VALUE certified raters.

In this study, the following research questions were addressed:

1) Among this group of raters, is there at least one rater exhibiting statistically significant differences in leniency/severity?
    a. If so, which raters exhibit leniency/severity effects?
2) Is there a group-level rater halo effect suggested by the absence of significant differences in the element difficulties?
    a. If so, which raters exhibit halo effects?
3) Is a group-level restriction of range indicated by the absence of significant differences in examinee abilities?
    a. If so, which raters exhibit restriction of range effects?
4) Overall, how many raters do not exhibit leniency/severity, halo effect, or restriction of range rater effects?

Chapter 2: Literature Review

Higher education institutions prominently use performance assessments for student assessment (Kuh et al., 2015). Worthen, White, Fan, and Sudweeks (1999) argued that performance assessments have become "a pervasive part of our culture" (p. 350). Moreover, Wolf (1994) asserted that performance assessments may be "the second most widely used measurement procedure, exceeded only by teacher-made achievement tests" (p. 4923). Many institutions independently develop and implement their own performance assessments to fit their programmatic needs. Additionally, the American Association of Colleges & Universities (AAC&U) has developed and advocated for several performance-based assessment systems, such as the Valid Assessment of Learning in Undergraduate Education (VALUE) rubrics and the VALUE Institute.

Performance assessments have grown in popularity due to claims that performance assessments have increased fidelity to real-world situations and allow for better evaluation of higher order thinking and learning, in comparison to selected-response assessments (Linn, Baker, & Dunbar, 1991; Stecher, 2014; Wiggins, 1991). Nonetheless, performance assessments have more intensive logistical and resource demands than selected-response assessments (Downing, 2006; Gronlund, 2003; Hardy, 1995; Linn et al., 1991; Madaus & Kellaghan, 1993). One such demand is the need for human raters to score performance assessments.

Performance assessments have additional psychometric challenges due to rater effects, or systematic differences in how raters judge students' performances or products (Myford & Wolfe, 2003). Rater effects can invalidate the inferences stakeholders want to make of performance assessment scores. High quality rubrics have been developed and

rater training for use of these rubrics are employed to limit the impact of rater effects. Unfortunately, these methods are usually incapable of preventing rater effects altogether. More techniques are necessary to combat rater effects in order to preserve the validity of performance assessment scores. Developing methods for selecting raters may be advantageous.

The purpose of this literature review is to 1) describe the advantages and disadvantages of performance assessments, 2) discuss the quality control methods already in place in the form of rubrics and rater training, and 3) review the most common rater effects, leading up to a discussion of 4) how already collected performance assessment scores can potentially be used to determine which raters to select for future rating sessions.

Performance Assessments

Performance assessments consist of a performance task and the scoring process (Khattri, Reeve, & Kane, 1998). A performance task typically requires ratees to carry out a process (i.e. presenting, playing a recital) or construct a product (i.e. writing an essay, producing a video). In the scoring process, a rater evaluates the process being carried out or the completed product (Johnson, Penny, & Gordon, 2009). For this reason, performance assessments are often called constructed-response assessments. Sometimes performance assessments are referred to as alternative assessments in contrast to selected-response assessments, which are the most common assessment type in higher education (Wiley & Haertel, 1996). A multiple-choice test is a common example of a selected response assessment format. Selected-response assessments typically ask students to select the best answer among several possible response options (Downing, 2006). Critics

of selected-response assessments argue that such assessments are decontextualized and lack the validity of true-to-life scenarios. On the other hand, performance assessments require engagement in a process or completion of a product; thus, they are often purported to have better fidelity to real-life situations. Some performance assessments have even been dubbed authentic assessments because they can simulate real-world situations and "involve the performance of tasks that are valued in their own right" (Archibald & Newman, 1988; Linn, Baker, & Dunbar, 1991, p. 15; Stecher, 2014; Wiggins, 1991). There is tension between advocates of performance assessments and those of selected-response assessment (Cizek, 1991a, 1991b; Wiggins, 1991, 1993). The use of each assessment format is debated for three reasons: the cost of implementing each assessment format, the cognitive levels each assessment format tends to be able to assess, and the psychometric properties of scores from each assessment format.

Additional resource and logistical concerns of performance assessments. Performance assessment tends to require considerably more resources than selected-response assessment (Downing, 2006; Gronlund, 2003; Linn et al., 1991; Madaus & Kellaghan, 1993). While both assessment formats require highly skilled test writers, test piloting and revising, and preliminary data collection for validity evidence; performance assessments have additional resource concerns due to the subjective nature of the scoring process (Welch, 2006). Consequently, resources and expertise need to be allocated to developing a quality scoring guide, most commonly in the form of a checklist or rubric (Johnson et al., 2009). Subsequently, raters need to be provided adequate training to properly apply the rubric scoring criteria to products or performances. Rubrics and rater training are integral components of performance assessment administration for ensuring

that scores are meaningful and useful representations of student ability (AERA, APA & NCME, 2014; Khattri et al., 1998; Stiggins, 1987). Moreover, the literature generally recommends that no less than two raters score each performance or product (Johnson et al., 2009). Therefore, substantial time and resources must be dedicated to administering a performance assessment considering that a strong scoring guide needs to be developed, raters must be trained, examinees need to complete the performance task(s), and raters need to complete the scoring process.

Another logistical concern of performance assessment is that they are limited in the breadth of content that can be covered and behavior sample size. Given the same amount of time, students are able to complete considerably more selected-response tasks than performance-based tasks (Downing, 2006; Gronlund, 2003; Linn et al., 1991). Consequently, more content can be assessed with selected-response assessments than with performance assessments and many more samples of students' behavior can be collected. Scores based on more samples of students' behaviors tend to provide more reliable representations of students' ability. Unfortunately, due to time constraints, educators usually cannot administer the amount of performance assessment tasks necessary to broadly cover a construct with a large sample of examinee' behaviors. Therefore, performance assessment scores are generally narrower in their interpretation (due to limited content representation) and inferences may be less dependable (due to limited samples of students' behavior).

Finally, performance assessments tend to be more expensive to develop, administer, and score than selected-response assessments (Hardy, 1995). Although estimates vary, the cost of administering performance assessments adequately is

substantial (Picus, Adamson, Montague, & Owens, 2010). As such, performance assessment users must account for the logistical and resources demands associated with performance assessments (Cizek, 1991b; Topol, Olson, & Roeber, 2010). However, performance assessment advocates argue that sustained costs can be justifiable, especially if the scores obtained are accurate indicators of higher-order cognitive abilities that match stated objectives, or the purpose of assessment (Hardy, 1996; Picus et al., 2010; Wiggins, 1993).

Higher-order knowledge best assessed by performance assessments.

Advocates of performance assessments claim that performance assessments are better able to measure students' higher-order knowledge, skills, and/or abilities (KSAs) than selected-response assessments (Chickering, 1999; Lane & Stone, 2006; Wiggins, 1991). Selected-response formats provide a legitimate means of measuring knowledge and can be designed to assess higher-order KSAs, like analysis and critical thinking (Cobb 1998; Downing, 2006; Haladyna, 2004). However, performance assessment formats are usually better able to tap into higher order KSAs because examinees are required to engage in a process or create a product, which can directly elicit higher order cognitive abilities if designed to do so (Lane & Stone, 2006; Linn et al., 1991; Wiggins, 1991). Some performance assessments have been dubbed authentic assessments because they are able to integrate the nuance of real-life context into the assessments (Linn et al., 1991). Thus, in addition to having knowledge about a construct, examinees must also be able to apply and implement that knowledge for performance assessment tasks, which is a dimension of cognitive ability that is difficult for selected-response formats to assess.

Many proponents of performance assessments advocate for them because they tend to be more direct measures of students' higher-level KSAs when compared to selected-response assessments (Lane & Stone, 2006; Resnick & Resnick, 1996). Nonetheless, a performance assessment does not necessarily elicit the desired higher-order KSAs, nor will the performance assessment inherently simulate the real-world context (Linn et al., 1991). As mentioned, performance assessments need to be designed with considerable resources and studious effort so that they would evoke the desired KSAs. Even well-developed performance assessments result in scores that can have serious psychometric challenges.

Additional psychometric challenges of performance assessment scores. Performance assessment scores typically have more psychometric concerns than scores from selected-response assessments (Gronlund, 2003). Selected-response assessments tend to have many more items sampling students' behavior than performance assessments. Gathering strong reliability evidence is more challenging for shorter assessments (Cronbach, 1990; Traub & Rowley, 1991). Moreover, assessing a construct with adequate breadth is essential for making valid inferences about students' KSAs, which is challenging for performance assessments as they tend to consist of smaller samples of behavior and content (AERA, APA & NCME, 2014). Consequently, performance assessment users must consider the strength of evidence supporting score generalization to the construct of interest (Haertel, 1999). While performance assessments can gauge the depth of understanding, the lack of breadth and behavior sampling results in limitations to the reliability and validity of interpretations made to a construct based on performance assessment scores (Messick, 1996).

Reliability and validity of performance assessment scores are further complicated by the manner of scoring in ways that selected-response assessments are not. Unlike selected-response assessments, performance assessments do not usually have an objectively clear correct or incorrect responses. Instead, scoring performance assessments is a more complex, subjective, rater-mediated process performed by human judges or sometimes by computer algorithms (Engelhard, 2002; Johnson et al., 2009). Additional error to scores can be introduced because of the subjective nature of the scoring process (Linn, 1993). Therefore, additional evidence must be presented for performance assessments to demonstrate that scores primarily indicate students' KSAs and not the rater (AERA, APA, & NCME, 2014). In order to have valid interpretations based on performance assessment scores, scores should not depend on the rater. As such, several ways of conceptualizing rater reliability are used by practitioners.

Reliability is most often operationalized as either consensus or consistency between raters (Stemler, 2004). High consensus reliability means that raters judging the same products or performances generally give the same scores. High consistency reliability means that raters judging the same products or performances generally rank-order student work in the same way. However, the information gathered from one type of reliability may contradict the other type. For instance, poor agreement between raters may be observed even when these raters exhibit high consistency across students (Eckes, 2015; Stemler, 2004). This can occur if a severe rater and lenient rater judge the same student work. Students would be rank-ordered similarly across raters, resulting in high consistency; however, there would be low agreement between these raters because their scores do not match one another. This contradictory reliability evidence may be

confusing for educators and stakeholders, especially since many researchers do not clearly provide a rationale for the type of rater reliability evidence they choose to use.

Assessment practitioners should not forgo performance assessments in favor of selected-response assessments by default due to the additional psychometric challenges and resources demands of performance assessments. Instead, decisions between a selected-response or performance assessment format should be based primarily on the purpose of the assessment and the logistical considerations (Lane & Stone, 2006; Schmeiser & Welch, 2006). As discussed, part of the purpose of an assessment has to do with the cognitive level and the depth and breadth of content coverage to be assessed as well as the inferences that will be made from assessment scores. Some logistical considerations involve accounting for the administration time available and additional resources that can be dedicated to the assessment. Performance assessments can be effectively used to garner information about higher-order KSAs if designed by a sound development process. Developing a strong scoring guide and providing rater training for use of this scoring guide are crucial steps in the sound development of a performance assessment that can produce strong psychometric evidence for score interpretations (Welch, 2006).

Rubrics and Rater Training

Rubrics are the most prominent scoring guides for scoring performance assessments (Saal, Downey, & Lahey, 1980). Rubrics are essential for producing performance assessment scores with adequate psychometric properties (Welch, 2006). Rubrics can be developed to be either holistic or analytic. Figure 1 displays the features of a rubric based on the example of part of AAC&U's Critical Thinking VALUE rubric

(Appendix A, p. 71). The number of elements and scoring criteria vary depending on the purpose of the performance assessment and the product or performance that will be assessed by the rubric. This rubric is an example of an analytic rubric because it contains more than one element. Analytic rubrics allow for various dimensions of a construct to be evaluated individually, generating multiple scores within the same assessment (Moskal, 2000; Welch, 2006). Contrastingly, a holistic rubric is designed to provide one score generated from an overall evaluation of the examinee's performance (Gronlund, 2003; Huot, 1990; Lane, 2014).

While neither rubric type is inherently better, the type of rubric that one develops and employs should be considered carefully. If the elements of an analytic rubric are not distinct from one another, then raters may be unable to differentiate between them — causing similar scores across rubric elements (DeCotiis, 1977; Johnson et al., 2009). On the other hand, a holistic rubric can be problematic if used on a construct where several different distinct elements are elicited. Raters may be confused on how to generate a single score when a product or process shows features of high and low performance across multiple criteria (Barkaoui, 2007). The type of rubric designed and employed depends on the theoretical framework underlying the construct being assessed and the kind of information one is interested in gathering from the assessment. Moreover, the choice between a holistic or analytic rubric may influence the psychometric quality of ratings (Lane & Stone, 2006; Wiggins, 1998).

Regardless of rubric type, scoring criteria should be developed for each element in such a way that raters are able to use the scoring criteria to differentiate examinees of varying abilities (Johnson et al., 2009). Furthermore, this scoring criteria should clearly

articulate the continuum of examinee ability underlying the dimension of the skill being measured (Wiggins, 1998). The ability continuum made explicit by the scoring criteria should have proficiency levels indicating the degree of skill represented by examinees' work. Similar to the choice between holistic and analytic rubric design, the number of proficiency levels depends on the theory underlying the construct and the type of information desired from the assessment. If the scoring criteria of the element is broken up into too many or too few proficiency levels, then differences between levels will be indistinguishable or muddied causing raters to be confused (Landy & Farr, 1980; Lane & Stone, 2006). Raters should be able to accurately separate and place students along the ability continuum into proficiency levels on each element.

Three recommendations guide the construction of rubrics (Tierney & Simon, 2004). First, the scoring criteria must clearly define the qualities at each score level. Second, the score criteria of each proficiency level should build upon the previous score. Third, the language used in scoring criteria across proficiency levels should be consistent. In other words, scoring criteria should grow in quantity, quality, or intensity across proficiency levels and new scoring criteria should not be introduced in subsequent proficiency levels within the same element (Popham, 1997; Wiggins, 1998). Furthermore, scoring criteria within proficiency levels should be presented descriptively, with behavioral anchors, rather than with subjective judgements (Moskal, 2000). For instance, descriptors such as "some" or "a lot" evoke subjective judgement of raters as to the meaning of "some" or "a lot," which can vary from rater to rater. If possible, a numerical description could be provided to anchor the meaning of "some or "a lot" in the proficiency levels of the scoring criteria. Clear descriptions of the scoring criteria at each

proficiency level helps raters accurately differentiate among students and assign appropriate scores for each element (Moskal & Leydens, 2000).

Rubrics are intended to guide raters through the scoring process by making explicit the attributes that are of most value in the performance task and operationalizing the different degrees of achievement (Lane & Stone, 2006). Rubrics aid in systematizing the method by which raters score performance assessments (Johnson et al., 2009; Tierney & Simon, 2004). The objective scoring structure provided by rubrics makes scoring performance assessments less subjective thereby improving score credibility and trustworthiness. The degree to which the scoring process is the same across raters strengthens the claim that scores represent examinee ability rather than rater effects (Stiggins, 1987). Rater training is another quality control mechanism usually employed to align raters to a rubric in order to increase the degree to which the scoring process is the same across raters.

Early rater training methods focused primarily on warning raters against common rater effects like leniency/severity, halo, and restriction of range. These methods have been shown to successfully reduce psychometric errors as defined by such rater effects (Bernardin & Walter, 1977; Borman, 1975; Borman, 1979; Ivancevich, 1979; Latham, Wexley, & Purcell, 1975). Nonetheless, researchers have contended that simply reducing psychometric error does not necessary translate to improved accuracy of ratings (Bernardin & Pence, 1980; Borman, 1975; Borman, 1979; Smith). As such, contemporary rater training methods focus more on familiarizing raters with the scoring criteria and how it should apply to an examinee's product or process (Bernardin & Buckley, 1981; Bernardin & Pence, 1980; Borman, 1979; Gordon, 1970). Aligning raters' interpretation

and application of the scoring criteria to the intended interpretation and application of the rubric should increase the degree to which the scoring process is the same across raters and produce consistent scores.

Nonetheless, the structure and practice of rater training programs vary and are met with varying degrees of success in terms of reducing the systematic errors of leniency/severity, halo, and restriction of range (Myford & Wolfe, 2003). At times, rater training only resulted in short-term improvements in psychometric score quality (Bernardin, 1978); while other studies found that only extensive training was effective in reducing rating errors (Bernardin & Walter, 1977; Brown, 1968; Latham, Wexley, & Pursell, 1975; Wexley, Sanders, & Yukl, 1973). Not only is more research needed into designing more effective rater training programs, but additional quality control measures are needed to improve the psychometric quality of scores and reduce the impact of rater effects on score validity.

Rater Effects

Although well-developed rubrics and rater training help structure a more objective scoring process, rater judgement continues to be an integral aspect of performance assessment ratings (Eckes, 2009; Myford & Wolfe, 2003). Scullen, Mount, and Goff (2000) defined rater effects as a "broad category of effects [resulting in] systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee" (p. 957). Performance assessment ratings are typically produced by using rater judgments thus they are considered "rater-mediated" (Engelhard, 2002). Ratings represent raters' perception of examinees work, raters' interpretations of the rubric, and raters' analysis of how examinees performance and the

rubric align. Ideally, all raters would interpret the scoring criteria as intended by rubric developers and apply them consistently across ratees. However, raters' interpretation of how the rubric should be applied to examinees work does not always align with the intended interpretation and use of the rubric. Leniency/severity, halo, and restriction of range are some of the most prominent, well researched rater effects that result from systematic variation due to rater mediation in the scoring process (Myford & Wolfe, 2003).

Leniency/severity. Leniency and severity are denoted by raters consistently assigning higher or lower scores, respectively, across ratees (Eckes, 2009, 2015; Engelhard, 1992; Saal et al., 1980). With respect to the average scores assigned by all raters, a severe rater consistently assigns lower scores on average across all ratees and a lenient rater consistently assigns higher scores on average across all ratees (Bond & Fox, 2015; Eckes, 2015; Wolfe, 2004). Ideally, all raters would interpret and apply rubric criteria in the same way resulting in similar average rating severity, implying that raters are interchangeable (Myford & Wolfe, 2004). In most research, interchangeable rater severity is assumed (Lunz, Wright, & Linacre, 1990). However, upon investigation raters usually exhibit significant differential severity from one another (Eckes, 2005; Han, 2015; Lunz, et al., 1990). Because rater mediated scores are used as a proxy for student ability, consistently severe scores underestimate student ability while consistently lenient scores overestimate student ability — both of which are problematic.

Traditionally, three methods have been used to identify if a leniency/severity effect is present among rater data (Saal, et al., 1980). Method one compares the mean ratings of each element with the midpoints of the proficiency levels. If the mean rating of

an element is considerably higher than the rating scale midpoint when the group of examinees have mean scores near the midpoint, then there may be evidence of leniency for that element. If the mean rating of an element is considerably lower than the rating scale midpoint, then there may be evidence of leniency for that element. The second method uses analysis of variance (ANOVA) to check for a statistically significant rater main effect. This is a G-theory method of variance components analysis (Brennan, 2001). If statistical significance is found for rater main effect, then there is evidence of leniency or severity depending on the direction of the main effect. In the third method, the degree of skewness in the frequency distributions of the ratings for each element are examined. When examine performance is not skewed, then a high degree of skewness indicates the presence of the leniency/severity rater effect; positive skew indicates rater leniency while negative skew indicates rater severity.

Halo. Halo is characterized by highly correlated scores across elements of a single ratee's product due to either (1) raters' inability to differentiate among distinct rubric elements (Borman, 1975; Saal et al., 1980); (2) raters allowing a general impression of the ratee impact scores for the distinct rubric elements (Schmidt & Hunter, 1980; Thorndike, 1920); or (3) raters allowing ratee's performances on an independent element impact scores on other distinct elements (Robbins, 1989). A halo effect can be problematic as it represents an inaccurate dependency among independent rubric elements that would stem from a holistic scoring schema rather than an analytic scoring schema (Engelhard, 1994). However, similar scores across elements that are correlated may be warranted and accurate if (1) rubric elements are not independent of one another (Bartlett, 1983; Cooper, 1981; Murphy, 1982; Pulakos, Schmitt, and Ostroff, 1986), (2)

students' abilities are actually similar across elements (Murphy & Cleveland, 1991; Solomonson & Lance, 1997), or (3) the rubric scoring criteria are not clearly differentiable (Nisbett & Wilson, 1977).

Traditionally, four distinctive methods have been used to identify if a halo effect is present among rater data (Saal, et al., 1980). Method one examines the intercorrelations among ratings on suspected elements. High correlations may suggest rater inability to discriminate among elements and therefore may be evidence of halo. The second method uses factor or principal-component analyses of the element intercorrelation matrix. If a few factors or principal components are found to explain a large part of score variance, then halo may be present in the ratings. In the third method, variances (or standard deviations) of each rater's scores of a particular ratee across all rubric elements are examined. Small standard deviation or variance estimates across the element scores are an indication of halo effect. For the fourth method, ANOVA is conducted, focusing on the rater by ratee interaction (Myford & Wolfe, 2004). This is a G-theory method of variance components analysis (Brennan, 2001). A statistically significant interaction lends evidence of halo effect, especially if the interaction explains a large portion of the variance in the ratings.

Restriction of Range. Restriction of range occurs when raters limit their judgements to a portion of the grading criteria or score levels and may be due to raters' inability to distinguish between scoring criteria across score levels (Myford & Wolfe, 2003). Restriction of range often manifests as either central tendency or extreme scoring. Central tendency is characterized by scores clustered around the midpoint of the scoring levels due to raters' avoidance of using extreme scoring levels (DeCotiis, 1977; Landy &

Farr, 1983; Long & Pang, 2015; Saal et al., 1980). Conversely, extreme scoring is characterized by scores clustered around either end of the scoring levels (note that patterns of extreme scoring limited to either the upper end of the scoring levels or the lower end cannot clearly be disentangled from rater severity or leniency, respectively). Limiting rater scores is problematic as lower quality products tend to be over-rated while higher quality products tend to be underrated; consequently, impeding the aim of normative assessment which is to separate ratees along a continuum of ability (Bandalos, 2018).

Central tendency is a specialized case of the restriction of range effect where the scores are clustered around the midpoint; however, the range of scores can be clustered elsewhere along the scoring levels (Saal et al., 1980). Restriction of range around the upper end of a scoring level can result from rater leniency while clustering around the lower end of a scoring level can stem from rater severity. Moreover, a halo effect consists of similar scores assigned across rubric elements resulting in a restriction of range at any score level. For instance, an examinee receiving a score of three on a five-point scale for each of three distinct elements would appear as both halo effect and central tendency. Thus, evaluating ratings for restriction of range is of utmost importance since many rater effects may manifest more broadly as a restriction of range effect (Engelhard, 1994).

Traditional evidence of central tendency in rating data stems from how close the average rating for an element is to the midpoint of the rating scale (DeCotiis, 1977; Landy and Farr, 1983). Traditionally, three methods have been used to identify if a restriction of range effect is present among rater data (Saal, et al., 1980). Method one examines the degree of kurtosis (i.e., peakedness) of the frequency distribution for the

scores on an element. A highly peaked distribution is indicative of restriction of range in the rating data. The second method conducts a rater by ratee by element ANOVA, focusing on the ratee main effect. If the ratee main effect is non-significant, then there is evidence of restriction of range because raters were not able to use the rating scale to discriminate between ratees in terms of their proficiency levels. This is a G-theory method of variance components analysis (Brennan, 2001). In the third method, the standard deviation of ratings across all ratees for an element are examined. The smaller the standard deviation, the greater the restriction-of-range effect.

Special consideration for leniency/severity, halo, and restriction of range rater effects should be made by using statistical modeling techniques when evaluating rater scores for accuracy.

Evaluating Scores for Rater Effects using MFRM

The Many-Facets Rasch Measurement (MFRM; Linacre, 1989) model estimates students' expected scores and has been proposed for the evaluation of rater effects in performance assessment scores (Eckes, 2015; Engelhard, 1992, 1994; Myford & Wolfe 2003). Rater effects research is limited by the fact that most accurate performance assessment scores are often unknown (Engelhard, 1996; Wolfe, 2004). Thus, statistical modeling techniques, like MFRM, enable researchers to estimate expected scores for each rating that can represent the most accurate score (e.g. Wolfe, 2004; Wu & Tan, 2016).

Various sources of variability believed to influence examinees' scores can be included in the MFRM model as facets (Eckes, 2009). Adding a rater facet to the MRFM produces estimates of the degree that rater effects impact examinee scores. Specifically,

with a rater facet, the MFRM can run statistical significance tests and produce effect size values regarding variability in rater leniency/severity or presence of restriction of range (Myford & Wolfe, 2004). Additionally, adding a rubric element facet can allow for the evaluation of how raters use the individual rubric elements. Specifically, following estimation of the MFRM with an element facet, researchers can evaluate statistical significance tests and effect size values regarding the variability across elements, indicating the presence of halo effect (Myford & Wolfe, 2004).

From the MFRM we can generates model-implied scores. Model-implied scores are estimated based on the facets specified in the model and are believed to be invariant across raters (Engelhard, 1992). In other words, a model-implied score represents the score an examinee ought to have received if scored by a rater of average leniency/severity. Examinees' model-implied scores are generated by accounting for how individual raters may have influenced examinees' scores (Stemler, 2004). A unique advantage of MFRM modeling is that the MFRM model can provide diagnostic information regarding which raters are showing evidence of which rater effects (Myford & Wolfe, 2003; Sudweeks, Reeve, & Bradshaw, 2005). This diagnostic information can be used to determine which raters need additional rater training or even which raters should be invited back for future ratings.

Study Purpose and Research Questions

The literature suggests that complete elimination of rater effects is unlikely, even with well-developed rubrics and strong rater training (Cronbach, 1990; Wu & Tan, 2016). As such, the purpose of the current study is to provide an additional quality control tool through the diagnostic evaluation of rater data to make recommendations for rater

selection. In order to do so, VALUE Institute scores are evaluated for rater effects. Of specific interest is the evaluation of diagnostic rater leniency/severity, halo effect, and restriction of range. Recommendations for rater selection can be made based on raters not exhibiting such rater effects. Moreover, it is important that VALUE rubric scores produced by the VALUE Institute are psychometrically sound and backed with evidence to support their interpretations and uses given that the VALUE Institute offers paid rating services using VALUE rubrics to higher education institutions so that these institutions can make institution-level inferences regarding students' abilities of higher order skills. Thus, this information is useful for the VALUE Institute and higher education institutions using their services, as it may have implications for the interpretations of VALUE Institute scores. This information can be particularly useful for the VALUE Institute as it can have implications for rater training and *selection* of VALUE certified raters.

In this study, the following research questions were addressed:

1) Among this group of raters, is there at least one rater exhibiting statistically significant differences in leniency/severity?
    a. If so, which raters exhibit leniency/severity effects?
2) Is there a group-level rater halo effect suggested by the absence of significant differences in the element difficulties?
    a. If so, which raters exhibit halo effects?
3) Is a group-level restriction of range indicated by the absence of significant differences in examinee abilities?
    a. If so, which raters exhibit restriction of range effects?

4) Overall, how many raters do not exhibit leniency/severity, halo effect, or restriction of range rater effects?

Chapter 3: Method

Participants

Ratee Participants. Data on all participants were collected by AAC&U VALUE Institute. Ratees consisted of students from two- and four- year colleges and universities from across the United States. Student work from various undergraduate credit levels were collected. Work samples consisted of but were not limited to essays and presentations. These work samples are sometimes referred to as artifacts. Data were collected from 6610 students, with 5138 from the Critical Thinking VALUE Rubric and 4290 from the Written Communication VALUE Rubric. Most examinees provided data for both rubrics.

Table 1 displays key demographic information for the ratee sample as a whole and by VALUE Rubric subsamples; however, several variables had a high degree of missing data, ranging up to 28% regarding Federal Pell Grant eligibility. Overall, demographic characteristics were similar for both ratee subsamples: 52% female and 32% male; 63% White, 10% Hispanic of Latino, 5% Black and 3% Asian; and 73% of ratees were from 19 to 24 years old. A quarter of ratees were eligible for the Federal Pell Grant, whereas 47% were not. Finally, most ratees attended a 4-year institution, 52% were in the public sector and 29% were in the private sector, and 17% attended a public 2-year university.

Raters. Two hundred and twenty-one raters were employed by the VALUE Institute to rate student work, with 118 raters for the Critical Thinking VALUE Rubric and 104 for the Written Communication VALUE Rubric. Only one rater scored work samples for both rubrics. Raters were recruited from a pool of higher education members

who self-selected into a VALUE rubric-calibration rater training program. Most raters were academic faculty. All raters were calibrated to the rubric(s) they were hired to rate artifacts with; however, raters' experience and use of VALUE rubrics varied.

Measures

Two VALUE rubrics were used to rate student work: Critical Thinking VALUE Rubric and Written Communication VALUE Rubric (See Appendix A and B for VALUE Rubrics). Each rubric is presented with a statement that briefly covers the design and purpose of the VALUE rubrics. The statement emphasizes that all VALUE rubrics were created by teams of faculty experts from various higher education institutions across the United States. For each VALUE rubric, the teams examined numerous campus rubrics and related documents to articulate fundamental scoring criteria with performance descriptors characterizing progressively more sophisticated levels of ability. These VALUE Institute rubrics were designed to assist the scoring process of various performance assessment tasks relating to each domain. They were intentionally designed to be flexible in order to meet the needs of educators on the individual level, such as for a particular program or institution (AAC&U, 2019). The utility of VALUE rubrics depends, in part, on assignment characteristics such as if all rating criteria are elicited by the prompt. Scores range from 0 to 4 on all elements across each VALUE rubric. Each VALUE rubric provides a definition for the domain the rubric is designed to assess as well as "Framing Language" or how the rubric is intended to be used. Additionally, most VALUE rubrics provide a glossary to clarify important terms used in the scoring criteria.

AAC&U investigated the validity of scores produced by VALUE rubrics by employing an argument-based approach (AAC&U, 2019; Kane, 2006). Kane (2001)

provided a strategy for developing an effective validity argument. First, the inferences and assumptions made in the interpretation of assessment scores must be explicated. Then, the robustness of the inferences and assumptions must be evaluated by all available validity evidence. This is known as the argument-based approach to validation. Strong validity arguments are backed by validity evidence. This validity evidence should satisfy the inferences and assumptions of assessment score interpretation and use.

The AAC&U validation effort based its argument-based framework on a revised version of Perie's (2013) interpretive argument for VALUE (see Appendix C). Perie's interpretive argument was specifically written for the VALUE rubrics to evaluate the degree to which and the conditions or assumptions that must be satisfied for the appropriate use of scores generated by VALUE rubrics. The interpretive argument consisted of 11 claims. However, the validation effort only focused on the six claims directly related to the VALUE rubrics (AAC&U, 2019). The assumptions of each claim were evaluated based on evidence from various sources, such as peer-reviewed journal articles and AAC&U-commissioned research. Evidence regarding each of the assumptions focused primarily on the development and design of VALUE rubrics, the calibration training given to VALUE Institute raters, how the VALUE rubrics are used, psychometric properties of data produced through VALUE rubric application, and the feedback of VALUE rubric users. Based on the strength of the validity argument, the validation team concluded with three strengths about the validity of VALUE rubrics. First, as intended and practiced, VALUE rubrics can be applied to numerous courses in a variety of disciplines. Second, the VALUE rubric rating scales appropriately distinguish among different levels of performance that faculty find relevant and understandable.

Third, trained faculty can use VALUE rubrics to evaluate student work and generate meaningful scores representative of student ability. Each VALUE rubric employed in the study is described further.

Critical Thinking VALUE Rubric. The Critical Thinking VALUE Rubric defined critical thinking and provides suggestions of student work that the rubric can be applied to: "Critical thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion" (Appendix A, p. 70). Note that critical thinking is defined as an investigative process of analysis that is transdisciplinary. While the Critical Thinking VALUE Rubric is designed to be used with many different assignment types, several recommendations are made regarding assignments that will extract the best information through the Critical Thinking VALUE Rubric:

> "**Critical thinking can be demonstrated in assignments that require students to complete analyses of text, data, or issues**. Assignments that cut across presentation mode might be especially useful in some fields. If insight into the process components of critical thinking (e.g., how information sources were evaluated regardless of whether they were included in the product) is important, assignments focused on student reflection might be especially illuminating." (Appendix A, p. 70)

Based on their experience scoring student work for the VALUE Institute, raters have provided recommendations for the assignment characteristics that are most assessable by the Critical Thinking VALUE Rubric (AAC&U, 2019). Assignments should require at least two viewpoints, including the student's opinion. For instance, students can analyze the positions of two different political parties and then present their own views for a civics assignment. Responses to assignments should be comprehensive (e.g. longer than

one page for an essay) and can be in the form of an evidence-based research paper with sources or a position paper that requires defending an argument and its conclusion. Five elements are designed to encompass the assessment of critical thinking.

Element A. Element A is labeled "Explanation of issues." In this element, students are rated on their ability to clearly and comprehensively describe an issue or problem that requires critical thinking. A low scoring artifact may state an issue or problem without providing enough information to convey the nuance of the situation that requires critical thinking. A high scoring artifact provides the relevant information necessary to understand the issue or problem without ambiguities in the description.

Element B. Element B is labeled "Evidence." In this element, students are rated on their ability to select and use information to investigate a point of view or conclusion. A low scoring artifact may consider expert opinions as facts and lack critical evaluation or interpretation of the information taken from sources. A high scoring artifact recognizes expert viewpoints as opinions and questions them appropriately. Furthermore, the information taken from sources are evaluated or interpreted into an appropriate and coherent viewpoint.

Element C. Element C is labeled "Influence of context and assumptions." In this element, students are rated on their ability to analyze how assumptions and context impact their position. A low scoring artifact may not recognize the presence of an assumption or miss important contextual considerations of the student's position. A high scoring artifact systematically and methodically analyzes the assumptions of the student's position and the assumptions others may hold. Furthermore, the relevance of contextual factors to the student's position is considered and their impact evaluated.

Element D. Element D is labeled "Student's position (perspective, thesis/hypothesis)." In this element, students are rated on the comprehensiveness of the position they present. A low scoring artifact may only state the specific position in a one-dimensional way. A high scoring artifact presents a specific position by taking into account the complexities of issues and acknowledges the limitations of the specific position as well as alternative viewpoints.

Element E. Element E is labeled "Conclusions and related outcomes (implications and consequences)." In this element, students are rated on their ability to logically evaluate evidence and perspectives to make appropriate conclusions and related outcomes. A low scoring artifact may oversimplify consequences and implications or selectively reference only evidence supporting the student's conclusions. A high scoring artifact evaluates a range of evidence, including opposing viewpoints, and presents them in a logical flow leading up to the conclusions and related outcomes.

Written Communication VALUE Rubric. The Written Communication VALUE Rubric defined written communication and provided guidelines for assignment characteristics that are important for alignment with the rubric: "Written communication is the development and expression of ideas in writing. Written communication involves learning to work in many genres and styles. It can involve working with many different writing technologies, and mixing texts, data, and images. Written communication abilities develop through iterative experiences across the curriculum" (Appendix B, p. 73). Note that the rubric defined written communication contextually, emphasizing the rhetorical nature of written communication skills. As such, several suggestions were made regarding the use of the Written Communication VALUE Rubric for assessment:

> "Evaluators using this rubric must have information about the assignments or purposes for writing guiding writers' work. Also recommended is including reflective work samples of collections of work that address such questions as: What decisions did the writer make about audience, purpose, and genre as s/he compiled the work in the portfolio? How are those choices evident in the writing -- in the content, organization and structure, reasoning, evidence, mechanical and surface conventions, and citational systems used in the writing? This will enable evaluators to have a clear sense of how writers understand the assignments and take it into consideration as they evaluate." (Appendix B, p. 73-74)

Based on their experience scoring student work for the VALUE Institute, raters have provided recommendations for the assignment characteristics that are most assessable by the Written Communication VALUE Rubric (AAC&U, 2019). Assignments should require high-quality writing and sources or citations. Responses can be in various forms, such as an evidence-based paper, a literary essay or analysis, an expository or persuasive essay, a lab report or a reflection. Five elements are designed to encompass the assessment of critical thinking.

Element A. Element A is labeled "Context of and Purpose for Writing." In this element, students are rated on their ability to consider the audience, purpose, and circumstances surrounding the writing task(s). A low scoring artifact may gloss over the context, audience, or purpose of the assigned task; perhaps limiting the audience to their instructor or themselves. A high scoring artifact clearly focuses all elements of the work around the context, audience, or purpose of the assigned task.

Element B. Element B is labeled "Content Development." In this element, students are rated on their ability to use content that is appropriate and relevant to the writing task(s). A low scoring artifact may only use appropriate and relevant content to

superficially develop ideas in a small section of the writing task. A high scoring artifact uses appropriate and relevant content that shapes the entire response and compellingly explores ideas within a subject to the point of mastery.

Element C. Element C is labeled "Genre and Disciplinary Conventions." In this element, students are rated on their ability to follow "formal and informal rules inherent in the expectations for writing in particular forms and/or academic fields" (Appendix B, p. 75-76). A low scoring artifact may only follow the appropriate expectations for basic organization, content, or presentation. A high scoring artifact exhaustively follows appropriate expectation of given a specific discipline or writing task(s), from organization and content to formatting and stylistic choices.

Element D. Element D is labeled "Sources and Evidence." In this element, students are rated on their ability to use appropriate, high-quality sources. A low scoring artifact may unsuccessfully attempt to reference sources to support ideas in the writing. A high scoring artifact develops ideas with sources that are credible and relevant to the discipline and genre of the writing.

Element E. Element E is labeled "Control of Syntax and Mechanics." In this element, students are rated on their ability to logically evaluate evidence and perspectives to make appropriate conclusions and related outcomes. A low scoring artifact will exhibit a high degree of errors in language usage that impedes meaning. A high scoring artifact is virtually error-free and communicates meaning to readers with clarity and fluency of language use.

Dependability of VALUE Rubric Scores. AAC&U investigated the interrater reliability of scores generated with the Critical Thinking and Written Communication

rubrics in 2015-2016 (AAC&U, 2019). Interrater reliability was evaluated using ordinal weights in four interrater reliability tests: (1) percent agreement, (2) Cohen's kappa, (3) Brennan-Prediger, and (4) Gwet's AC coefficients (Gwet, 2010). Percent agreement examines the portion of raters who generate the same score. Cohen's kappa takes chance agreement into account in the same way as a chi-square test of independence where raters are assumed to be independent. However, Cohen's kappa is limited as it tends to be highly influenced by the marginal distribution. Brennan-Prediger accounts for chance by adjusting for the number of proficiency levels in the rubric (Gwet, 2010). Gwet's AC adjusts for chance further by accounting for how hard it is for raters to rate an artifact. An artifact that is difficult for raters to judge will tend to have a uniform distribution of scores whereas an artifact that is easy to score will have ratings placed into the same proficiency level. For both rubrics, interrater reliability was moderate to strong according to most metrics: ranging from 88% to 94% according to the weighted percent of exact agreement, weighted Brennan-Prediger values from.56 to .77, and weighted Gwet's AC2 values from .60 to .84. However, Cohen's kappa was lower with values ranging between .26 to .39, likely due to the limitation of Cohen's kappa mentioned above (See Table 2).

Procedure

VALUE Rubric essay collection. All data were collected by the AAC&U VALUE Institute. Prior to collecting artifacts, VALUE representatives and higher education clients met to discuss client's assessment goals. Then, the VALUE institute provided guidelines for gathering a representative sample of student work that matched the established purpose (AAC&U, "Guide to Developing Your Sampling Plan"). Non-restrictive guidelines were provided for determining appropriate artifacts — primarily to

ensure that assignments align with VALUE rubric(s) and the assessment purpose. Artifacts were collected over the course of one academic year. Artifacts were scored shortly after each academic year. A variety of work samples were collected; however, artifacts were mostly essays and presentations.

Rating Process. The rating process occurred over the summer after the 2018-2019 academic year. Raters were recruited from a pool of individuals who self-selected into a VALUE rubric-calibration rater training program (AAC&U, 2019; S. Tang, personal communication, November 11, 2020). These are typically higher education members seeking professional development in how to apply VALUE rubrics to assess student learning of higher-order skills. Training consisted of interactive videos describing how to apply the VALUE rubric to student work in the scoring process — participants had an opportunity to discuss score discrepancies with a VALUE Institute member after scoring an artifact and submitting their scores for review. Each training session targeted a specific VALUE rubric (e.g. Critical Thinking, Written Communication, etc.). Clients signed up for the VALUE rubric on which they would like to be trained.

Subsequently, a VALUE Institute member contacted individuals in the training program to recruit them for VALUE Institute scoring. The individuals were asked to complete ratings for the VALUE Institute, primarily as a professional development experience. However, a small financial incentive was also provided. Ideally, raters were selected based on how closely they matched training artifacts to the scores determined by VALUE Institute members. Nonetheless, raters were typically selected based on their availability. Upon successfully completing the VALUE rubric-calibration rater training program, raters were designated as VALUE-certified raters and were eligible to rate

artifacts submitted to the VALUE Institute. All artifacts were de-identified prior to rating. Each artifact was scored by at least two trained, VALUE-certified raters.

Data Analysis

Data were received pre-screened by VALUE Institute. All students were scored by at least two raters. All data preparation was conducted using Excel and SAS Software Version 9.4, unless otherwise stated. All data analysis was conducted using FACETS, unless otherwise stated (Linacre, 2017b). Data analysis for the assumptions and research questions were addressed for the Critical Thinking VALUE rubric and the Written Communication VALUE rubric separately.

Data preparation. Only relevant data were extracted from the dataset received from VALUE Institute: student id, rater id, and corresponding ratings. During data screening, 248 cases were deleted for missing a student id and 15 cases were deleted for missing a rater id. No missing scores were found; however, values of zero were recoded as missing as was practiced in similar MFRM analysis of AAC&U VALUE Rubric data and related research of AAC&U VALUE Rubric data (Gregg, 2018; Hathcoat, 2018). This decision was made because VALUE Institute raters could assign values of zero, "representing an absence of evidence of student learning for that specific criterion" even though the VALUE rubrics consist of proficiency levels only ranging from one to four (AAC&U, 2017, p. 32). However, this absence of evidence could be due to a lack of student ability or because the assignment did not illicit skills for this criterion (Gregg, 2018; Hathcoat, 2018). In addition to the ambiguous meaning, the inclusion of zeros caused problems in MFRM modeling, which were remedied once zeros were removed by being coded as missing.

In total, there were 9,428 artifacts for analysis, with 5,138 artifacts for the Critical Thinking VALUE Rubric and 4,290 artifacts for the Written Communication VALUE Rubric. A master list of raters across all rubric data was created. Results of the present study were not linked directly to VALUE Institute raters. Raters will not be identified by name and they will be referred to by the rater id assigned for this study (as "rater 1," "rater 2," "rater 3," and so on) in all results.

Per the requirements of the FACETS software, student id and rater id were recoded to be sequential, starting from one. Scores were already in integer form, which is required by FACETS, and ranged from 0 to 4. Then, data were organized to meet FACETS specifications and exported as an Excel file.

Many-Facets Rash Measurement. The Many-Facets Rasch Measurement (MRFM) model was used to evaluate all research questions (Linacre, 1989). The MFRM model is an extension of the single-facet rating scale model (Andrich, 1978) and single-facet partial-credit model (Masters, 1982). These are expressions for testing situations where examinees can either get an item right or wrong. A dichotomous Rasch model including the facets of student and item can be defined as

$$ln\frac{P_{ni}}{1-P_{ni}} = \theta_n - \delta_i \quad (1)$$

where P_{ni} is the probability of student *n* answering *i* correctly,

θ_n is the ability of student *n*,

δ_i is the difficulty of item *i* (See Appendix D for a list of all equations).

However, performance assessments are rarely scored just as right or wrong. Instead, performance assessments use proficiency levels to represent degree of correctness. The

MFRM allows for multiple facets of polytomous-scored assessment items to be evaluated, such as with the VALUE rubric. All VALUE rubrics had 4 proficiency levels, ranging from one to four. Therefore, instead of estimating an examinee's probability of answering an item right or wrong, polytomous Rasch models include a rubric element facet. In so doing, polytomous Rasch models estimate an examinee's probability of receiving a given proficiency level as compared to the next lowest proficiency level. A polytomous Rasch model including the facets of student and rubric element can be defined as

$$ln\frac{P_{nik}}{P_{nik-1}} = \theta_n - \delta_i - \tau_k \quad (2)$$

where P_{nik} is the probability of student *n* being rated *k* on element *i*,

$P_{nik-1}$is the probability of student *n* being rated *k*-1 on element *ij*,

θ_n is the ability of student *n*,

δ_i is the difficulty of VALUE rubric element *i*,

and τ_k is the difficulty of score level *k* compared to score level *k*-1.

Moreover, a rater facets can be added since different judges evaluate performance assessment artifacts. Comparisons can be made across facets because all facets are placed on the same log odds (or logit) measurement scale (Bond & Fox, 2015). Model 1, a rating scale model, including the facets of student, rater, and rubric element can be defined as

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

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*,

θ_n is the ability of student *n*,

δ_i is the difficulty of VALUE rubric element *i*,

α_j is the severity of rater *j*,

and τ_k is the difficulty of score level *k* compared to score level *k*-1 (Eckes, 2015). An assumption is made when using the rating scale model that all raters used the set of rubric elements in the same way. Additionally, all rubric elements must be designed with the same number of proficiency levels to fit the requirement of the rating scale model (Bond & Fox, 2015; Myford & Wolfe, 2003). However, if the rubric elements are assumed to be used in their own individual ways, then a partial credit model can be specified, where proficiency levels vary by rubric element, Model 2, which can be defined as

$$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_{ik} \quad (4)$$

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*,

θ_n is the ability of student *n*,

δ_i is the difficulty of VALUE rubric element *i*,

α_j is the severity of rater *j*,

and τ_{ik} is the difficulty of score level *k* compared to score level *k*-1 for VALUE rubric element *i* (Eckes, 2015).

This partial credit model is more complex than the rating scale model because it estimates additional parameters for rubric element thresholds (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003). Nonetheless, for both MFRM models, the log-odds of students

obtaining scores of *k* are a function of the additive effects of their abilities (θ), the difficulty of the VALUE rubric element (δ), rater severity (α), and the difficulty of scoring in score level k compared to k-1 (τ; Eckes, 2009, 2015; Linacre, 2017a; Myford & Wolfe, 2003). Another partial credit model, Model 3, can be used by allowing the proficiency levels to vary by rater instead of varying by element as in Model 2. Model 3, including the facets of student, rater, and rubric element, can be defined as

$$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_{jk} \quad (5)$$

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*,

θ_n is the ability of student *n*,

δ_i is the difficulty of VALUE rubric element *i*,

α_j is the severity of rater *j*,

and τ_{jk} is the difficulty of score level *k* compared to score level *k*-1 for rater *j* (Eckes, 2015).

In this study, Model 1 was used for most research questions (1, 1a, 2, and 3) whereas Model 2 and Model 3 were used to evaluate research questions 2a and 3a, respectively. Research question 4 was evaluated based on the results of previous research questions and their corresponding analysis.

Joint-maximum likelihood estimation was used to estimate all MFRM models in FACETS 3.80.0 (Linacre, 2017b). Indices that are commonly used in the literature to evaluate rater-mediated scores for rater effects were used to evaluate each research question (e.g. Engelhard, 1992, 1994; Eckes, 2005; Wu & Tan, 2016). Although each

index is provided in output created by FACETS (Linacre, 2017b), an overview of each metric and its computation are provided for the benefit of the reader. Where appropriate, interpretations and ideal results for each metric are provided for each research question.

Fixed-effect chi-square. The fixed-effect chi-square is a significance test. It tests the null hypothesis of no differences in the logit values for a facet of measurement (e.g. student, rater, VALUE Rubric element), controlling for measurement error (Eckes, 2015; Myford & Wolfe, 2003). For instance, a non-significant chi-square for raters suggests that all raters exhibit the same severity, after controlling for measurement error. Rater is the facet of measurement in this study. The fixed-effect chi-square is defined as

$$x^2 = \sum (W_o * D_o^2) - \frac{(\sum W_o * D_o)^2}{\sum W_o} \quad (6)$$

where D_o is the estimated logit of the facet of measurement (leniency/severity of rater)

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 facet of measurement (Myford & Wolfe, 2003). However, like any statistical significance test, the fixed-effect chi square is sensitive to sample size. Consequently, even small differences in the logits of the facet of measurement can produce statistically significant fixed-effect chi square results in large samples (Eckes, 2015). The fixed-effect chi-square significance test, of the corresponding facet of measurement, will be used to evaluate research questions 1 (rater facet), 2 (element facet), and 3 (examinee facet).

Separation ratio. The separation ratio (G_o) is a measure of the spread of the logits associated with the facet of measurement relative to their precision (Eckes, 2015;

Myford & Wolfe, 2003). In other words, the separation ratio indicates the precision of the facet of measurement in spreading across the logit continuum. In order to calculate the separation ratio, the *true* standard deviation needs to be computed, defined as

$$SD_t^2 = SD_o^2 - MSE \tag{7}$$

where SD_o^2 is the observed logits' standard deviation of a facet of measurement and MSE is the average measurement error associated with that facet of measurement (Eckes, 2015).

Using the *true* standard deviation, the separation ratio can be defined as

$$G_o = \sqrt{\frac{SD_t^2}{MSE}} \tag{8}$$

The separation ratio ranges from zero to positive infinity, where values closer to zero indicate less spread of the facet of measurement across the logit continuum as compared to higher values (Eckes, 2015; Myford & Wolfe, 2003). Subsequently, G_o is used to calculate a separation index and reliability of separation.

Separation index. The separation index (H_o) indicates the number of different levels of the facet of measurement that are statistically significant (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003). The separation index is defined as

$$H_o = \frac{4\sqrt{\frac{SD_t^2}{MSE}} + 1}{3} \tag{9}$$

The separation index ranges from zero to positive infinity. So, if $H_o = 4.2$ for the rater facet, then the separation index suggests four distinct levels of raters "— that is, the spread of the rater severity measures is considerably greater than the precision of those measures" (Myford & Wolfe, 2004, p. 196). Ideally, H_o would be near 1.0 suggesting

that there is only one strata of raters, which would support the interchangeability of raters.

Reliability of separation. The reliability of separation (R_o) is analogous to traditional reliability indices (i.e. Cronbach's alpha) and ranges from zero to one (Myford & Wolfe, 2003). The reliability of separation estimates how reliably the facet of measurement can be separated along the logit continuum, where higher values are indicative of more reliable separation in the facet of measurement than lower values (Bond & Fox, 2015; Eckes, 2015). The reliability of separation is defined as

$$R_o = \frac{\frac{SD_t^2}{MSE}}{1 + \frac{SD_t^2}{MSE}} \tag{10}$$

and can be interpreted as the proportion of observed score variability in the facet of measurement that is not due to measurement error (Eckes, 2015). Essentially, reliability is simply true-score variance over true-score variance and error variance:

$$R_o = \frac{\frac{SD_t^2}{MSE}}{1 + \frac{SD_t^2}{MSE}} = \frac{\frac{SD_t^2}{MSE}}{\frac{MSE + SD_t^2}{MSE}} = \frac{SD_t^2}{SD_t^2 + MSE} \tag{11}$$

Reliability indices are often use for quantifying the magnitude of person separation. In that context, high reliability of separation is desirable. But for raters, "in many situations, the most desirable result is to have a reliability of rater separation close to zero, which would suggest that the raters were interchangeable, exercising very similar levels of severity" (Myford & Wolfe, 2004, p. 196). The separation index and the reliability of separation are used to evaluate research questions rather than directly using the separation ratio. According to Myford and Wolfe (2003), for raters, "the reliability of separation

index reflects potentially unwanted variation between raters in the levels of leniency/severity they exercised, that is, how different the rater severity measures are. (This is in direct contrast to interrater reliability, an index of how similar the rater severity measures are.) If one's goal is to have raters use one or more rating scales in a similar fashion, then low rater separation reliability is desirable" (p. 411).

Evaluation of MFRM assumptions. Local independence, unidimensionality, and correct model form are three MFRM assumptions that needed to be evaluated prior to data analysis.

Local independence. Local independence is satisfied if item responses are independent from one another after controlling for the construct of interest (DeMars, 2010). A violation of local independence indicates that the item is measuring a secondary construct or that the response of an item influences the response of another item (Marais & Andrich, 2008). Violations of local independence are problematic as they can influence parameter estimates (Li, Li, & Wang, 2010; Smith, 2005) and can inflate reliability estimates (Marais & Andrich, 2008; Wainer & Thissen, 1996; Wang & Wilson, 2005). One method to deal with violations of local independence is to sum the dependent items and treat them as a single polytomous item (DeMars, 2010; Marais & Andrich, 2008; Stone & Zhu, 2015).

Local independence would be met in this study if students' probabilities of receiving a particular score on a VALUE Rubric element were not related to a score they received on another element, after controlling for students' ability on the construct being measured. The assumption of local independence was evaluated in this study using Yen's Q3 correlations between residuals. Yen's Q3 values were adjusted for the mean Q3 and

compared to a critical value of .20 (Christensen, Makransky, & Horton, 2017; C. DeMars, personal communication, February 11, 2021). In this study, the assumption of local independence was considered satisfied for adjusted Yen's Q3 values not exceeding the .20 cutoff.

Unidimensionality. Unidimensionality means that all assessment items are assumed to measure only the one, common construct (Bandalos, 2018; DeMars, 2010). The assumption of unidimensionality was evaluated in this study by performing a Principal Components Analysis (PCA) on the standardized residuals. IBM SPSS Version 24 was used to perform the PCA. The following formula was used to estimate standardized residuals:

$$Z_{nij} = \frac{x_{nij} - e_{nij}}{\sqrt{w_{nij}}} \tag{12}$$

where x_{nij} represents 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} represents the model variance or the variability of the observed rating around its expected rating (Eckes, 2015).

The expected rating can be defined further as

$$e_{nij} = \sum_{k=0}^{m} kp_{nijk} \tag{13}$$

where *k* is a rating and p_{nijk} is the probability of student *n* obtaining score *k* on element *i* from rater *j*, given the model (Eckes, 2015).

The model variance can be defined further as

$$w_{nij} = \sum_{k=0}^{m} (k - e_{nij})^2 p_{nijk} \quad (14)$$

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

In the Rasch framework, PCA analyses are used to evaluate if there are systematic patterns in the residuals ("Dimensionality: Contrasts and Variances," n.d.). Such patterns among the residuals can indicate the presence of a secondary dimension, which is often called a contrast. The first dimension is removed by calculating the residuals, so the first contrast represents the second dimension. Thus, the PCA tests if any elements group on secondary contrasts. Each contrast can be represented by an eigenvalue that indicates the number of elements making up the contrast. Secondary contrast eigenvalues less than 2.0 suggest that less than two elements group on the secondary dimension. In this study, the assumption of unidimensionality was considered satisfied if the eigenvalues for the secondary contrasts were less than 2.0.

Correct model form. Correct model form refers to the assumption that the model used to analyze the data is fitting or appropriate. While data will never fit any model perfectly (Linacre, 2003), fit indices can be used to evaluate if the data fit the specified model well enough to provide useful estimates for answering the research questions. The assumption of correct model form was evaluated in this study by evaluating overall model fit and rater fit.

Overall model fit. The absolute value of the standardized residuals were examined to evaluate overall model fit. Standardized residuals represent the number of standard deviations an observed score deviated from the expected score. As such, standardized residuals of |2.0| indicate that the observed score deviated by two standard deviations from the expected score. Thus, standardized residuals greater than |2.0| suggest highly unexpected scores, because they are expected to appear less than 5% of the time in data that are consistent with the specified MFRM model (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003; Wright & Masters, 1982). Consequently, data were determined to fit the specified model well, overall, if less than 5% of the standardized residuals exceeded or were equal to |2.0|. If overall model fit was not satisfied according to this metric, then the sources of misfit would be investigated, and decisions would be made about excluding problematic raters.

Rater fit. Rater fit was evaluated because raters are the primary focus of analysis in this study. Rater fit was evaluated using Mean Square outfit (or unweighted mean squares) and Mean Square infit (or weighted mean squares). Mean Square outfit is defined as

$$MS_{U_j} = \frac{\sum_{n=1}^{N}\sum_{j=1}^{I} z_{nij}^2}{NI} \tag{15}$$

where *N* is the number of students the rater rated, and

I is the number of elements (Eckes, 2015).

The Mean Square outfit is simply the average of raters' squared standardized residuals (equation 9) for all students and elements. Mean Square infit values are weighted by statistical information, such that, ratings assigned in proficiency levels farther from

examinees' ability are weighted less heavily than ratings assigned to the closer proficiency levels because these extreme scores contribute less information to the model (Bond & Fox, 2015; Eckes, 2015). Mean Square infit is defined as

$$MS_{W_j} = \frac{\sum_{n=1}^{N}\sum_{j=1}^{I} z_{nij}^2 w_{nij}}{\sum_{n=1}^{N}\sum_{j=1}^{I} w_{nij}} \quad (16)$$

where all terms are defined as in equation 9 (Eckes, 2015).

Infit and outfit range from zero to positive infinity, with a value of 1.0 indicating perfect fit of rater scores to the model (Linacre, 2003). Overfit occurs with values less than 1.0, which suggests that observed ratings are more similar to ratings expected by the model than would be predicted by the model (Eckes, 2015; Linacre, 2003). Underfit occurs with values more than 1.0, which suggests that observed ratings are less similar to ratings expected by the model than would be predicted by the model. Infit and outfit are measures of effect size. However, both fit statistics can be transformed to a *t*-distribution to test the statistical significance of perfect model-data fit (Eckes, 2015). Nonetheless, using these metrics as indicators of both, effect size and statistical significance, is uncommon in Rasch measurement (DeMars, 2010). Consequently, infit and outfit were used as untransformed measures of effect size in this study.

Several similar benchmarks of acceptable rater fit based on infit and outfit have been proposed by Rash measurement experts. Linacre (2003) suggested that infit and outfit values between 0.5 and 1.5 indicate acceptable rater fit, while Bond and Fox (2015) proposed a narrower range of 0.7 to 1.3 as more appropriate for higher stakes assessment. While there are no strict benchmarks for acceptable infit and outfit values, values greater

than 2.0 indicate major distortions in model fit (Eckes, 2015; Linacre, 2003). Since the purposes of VALUE rubric scores are relatively low stakes, infit and outfit values between 0.5 and 1.5 were considered acceptable and values greater than 2.0 were flagged as indicators of major rater misfit.

After assumptions were tested, data were analyzed to evaluate each research question. In all analyses, facets were oriented so that higher logit values indicated more presence of that facet. In other words, higher logit values for the examinee facet represented more ability than lower logit values, higher logit values for the rater facet represented more severity in rating than lower logit values, and higher logit values for the element facet represented a more difficult element. The average examinee ability logit was freely estimated while the average logits of the rater and element facets were fixed to zero. The data analysis procedures and metrics used to evaluate each research question is described next.

Research Questions

Table 2 summarizes how each research question was evaluated, specifying the model that was estimated, the facet of interest, and which rater effect indicators were examined, along with a brief rationale for how the indicators relate to the research questions.

Research question 1: Among this group of raters, is there at least one rater exhibiting statistically significant differences in leniency/severity? Model 1 (equation 3), specifically the rater facet (α_j), was used to evaluate this research question. First, the fixed-effect chi-square (equation 6) of the rater facet was evaluated as an overall test of whether leniency/severity differed across raters. The null hypothesis stated that there was

no difference in rater severity, after controlling for measurement error. A statistically significant chi-square ($p < .05$) indicates that at least two raters have statistically significantly different leniency/severity logit scores (Myford & Wolfe, 2004).

Next, rater separation index and reliability of rater separation were evaluated with raters as the object of measurement. For the rater separation ratio (equation 8), the true standard deviation (equation 7) was computed using the observed standard deviation of the rater logits and the standard error associated with the rater logits. Moreover, the rater separation index (equation 9) was estimated to determine how many levels of rater leniency/severity were statistically significantly different (Myford & Wolfe, 2003). Ideally, the rater separation index will be low to suggest a few statistically distinct levels of rater leniency/severity as compared to larger values (Myford & Wolfe, 2004). Additionally, the reliability of separation (equation 10) for raters was estimated to indicate how reliably raters could be separated along the leniency/severity continuum (Myford & Wolfe, 2003). Ideally, the rater reliability of separation will be low to indicate that raters cannot be reliably separated along the leniency/severity continuum due to a high degree of similarity in leniency/severity (Myford & Wolfe, 2003; Myford & Wolfe, 2004).

Research question 1a: Which raters exhibit leniency/severity effects? Model 1 (equation 3), specifically the rater facet (α_j), was used to evaluate this research question. Individual raters' severity/leniency logit values were evaluated by visually inspecting a Wright map, also called a variable map or vertical ruler (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2004). The Wright map provided a visual representation of raters' leniency/severity; rank-ordering raters by their leniency/severity logit values. Ideally,

raters will be clustered near a logit value of 0.0 on the Wright Map, which would mean that raters are near average leniency/severity. If raters are spread across the logit continuum, then this would indicate that raters differ in their leniency/severity. Raters who were higher than 0.0 on the Wright map were considered to be more severe than the average rater (Bond & Fox, 2015; Eckes, 2015; Linacre, 2017a; Myford & Wolfe, 2004). Conversely, raters who were lower than 0.0 on the Wright map were considered to be more lenient than the average rater.

Next, for raters that visually appear to deviate on the Wright map, rater "fair averages" were examined. A fair average is the average expected rating for each rater based on the MFRM — a rater's average adjusted for the deviation of the ratees in each rater's sample from the overall ratee average across all raters and elements (Myford & Wolfe, 2004). Ideally, raters will have a similar observed average and model expected fair average. Finally, for the raters still suspected of exhibiting rater severity/leniency effect, the frequency counts were examined to confirm how the rater assigned scores. Raters that showed evidence of a rater severity/leniency effect on based on the spread of severity/leniency logits on the Wright map, extremely discrepant fair averages from observed average, and/or frequency analyses were determined to be exhibiting rater severity/leniency effect. A total count of such raters was recorded.

Research question 2: Is there a group-level rater halo effect suggested by the absence of significant differences in the element difficulties? In other words, are the raters, overall, distinguishing among the elements? Model 1 (equation 3), specifically the element facet (δ_i), was used to evaluate this research question. First, the fixed-effect chi-square (equation 6) of the element facet was evaluated as an overall test of whether

elements differed in difficulty. The null hypothesis stated that there was no difference in element difficulty, after controlling for measurement error. A statistically significant chi-square ($p < .05$) indicates that at least two elements have statistically significantly different difficulty logit values (Myford & Wolfe, 2004). If element difficulty is indistinguishable as would be indicated with a non-significant fixed-effect chi-square, then it suggests that raters assigned similar scores across elements. This could be due to a halo effect impacting raters' scoring process. As research indicates, VALUE rubric elements' difficulty should vary across the logit continuum, which would mean that certain elements are more difficult than others. Consequently, a significant chi-square test would produce evidence that a halo effect is not present (Myford & Wolfe, 2004).

Next, the element separation index and reliability of element separation were evaluated with element as the object of measurement. For the element separation ratio (equation 8), the true standard deviation (equation 7) was computed using the observed standard deviation of the element logits and the standard error associated with the element logits. Moreover, the element separation index (equation 9) was estimated to determine how many levels of element difficulty are statistically significantly different (Myford & Wolfe, 2003). Ideally, the element separation index will be higher to suggest more statistically distinct levels of element difficulty as compared to lower values (Myford & Wolfe, 2004). A group-level halo effect is more likely when no distinct levels of element difficulty are present. Additionally, the reliability of separation (equation 10) for elements was estimated to indicate how reliably raters can distinguish among elements (Myford & Wolfe, 2003). Ideally, the element reliability of separation will be higher to indicate that raters can reliably distinguish among elements due to a higher

degree of variation in element difficulty (Myford & Wolfe, 2003; Myford & Wolfe, 2004). A low element reliability of separation value can be due to a halo effect.

Research question 2a: Which raters exhibit halo effects? Results from Model 2 (equation 4), specifically the threshold by element facet (τ_{ik}), were used to evaluate this research question. Mean Square outfit (equation 14) and Mean Square infit (equation 15) were evaluated to determine if specific raters exhibited halo effect (Myford & Wolfe, 2004). If there is evidence that element difficulty varied, then raters exhibiting halo effect will be flagged with significantly higher infit and outfit mean-squares indices (values greater than 1.5). If there is evidence that element difficulty did not vary, then raters exhibiting halo effect will be flagged with significantly lower infit and outfit mean-squares indices (values less than 0.5). This would suggest that the rater was not able to differentiate reliably between conceptually distinct traits.

Next, for raters flagged for extreme infit and outfit mean-squares values, the number of times the rater assigned the same scores throughout elements was calculated. Ideally, there will be few instances that the rater assigned identical ratings across elements for elements with varying difficulty. Finally, for the raters still suspected of exhibiting halo effects, the frequency counts were examined to confirm how the rater assigned scores. Raters that showed evidence of halo effects on based extreme rater infit and outfit values, assigning the same scores throughout elements of varying difficulty, and/or frequency analyses were determined to be exhibiting halo effects. A total count of such raters was recorded.

Research question 3: Is a group-level restriction of range indicated by the absence of significant differences in examinee abilities? Model 1 (equation 3),

specifically the element facet (θ_n), was used to evaluate this research question. First, the fixed-effect chi-square (equation 6) of the ratee facet was evaluated as an overall test of whether ratee ability differed according to their logit scores. The null hypothesis stated that there was no difference in ratee ability, after controlling for measurement error. A statistically significant chi-square ($p < .05$) indicates that at least two ratees have statistically significantly different ability logit scores (Myford & Wolfe, 2004). If ratee ability is indistinguishable as would be indicated with a non-significant fixed-effect chi-square, then it suggests that raters assigned similar scores to ratees. This could be due to a restriction of range effect impacting raters' scoring process. Ideally, ratees will be distributed across the logit continuum, which would represent ratees differing in their ability estimates. Consequently, the chi-square test will be significant to produce evidence that a restriction of range effect is not present (Myford & Wolfe, 2004).

Next, the ratee separation index and reliability of ratee separation were evaluated with ratees as the object of measurement. For the ratee separation ratio (equation 8), the true standard deviation (equation 7) was computed using the observed standard deviation of the ratee ability logits and the standard error associated with the ratee ability logits. Moreover, the ratee separation index (equation 9) was estimated to determine how many levels of ratee ability are statistically significantly different (Bond & Fox, 2015; Eckes, 2015; Myford & Wolfe, 2003). Ideally, the ratee separation index will be large to suggest more statistically distinct levels of ratee ability as compared to smaller values (Myford & Wolfe, 2004). A group-level restriction of range effect is unlikely when distinct levels of ratee ability are present. Additionally, the reliability of separation (equation 10) for ratees was estimated to indicate how reliably ratees can be separated along the ability

continuum (Myford & Wolfe, 2003). Ideally, the ratee reliability of separation will be high to indicate that ratees can be reliably separated along the ability continuum due to a high degree of variation in their estimated ability logits (Myford & Wolfe, 2003; Myford & Wolfe, 2004).

Research question 3a: Which raters exhibit restriction of range effects?

Results from Model 3 (equation 5), specifically the thresholds by rater facet (τ_{jk}), were used to evaluate this research question. Mean Square outfit (equation 14) and Mean Square infit (equation 15) were evaluated to determine if specific raters exhibited restriction of range effects (Myford & Wolfe, 2004). Raters potentially exhibiting restriction of range effects will be flagged with significantly higher infit and outfit mean-squares indices (values greater than 1.5) or significantly lower infit and outfit mean-squares indices (values less than 0.5).

Next, for raters flagged for extreme infit and outfit mean-squares values, rating scale category thresholds and their outfit mean-square indices were evaluated to determine if poor rater fit was due to restriction of range (Myford & Wolfe, 2004). A rating scale category threshold indicates the logit value where the probability curves of two adjacent scale categories cross. In other words, a rating scale category threshold represents the point where an examinee has a 50% probability of being rated in either of the adjacent categories, as expected by the model. Rating scale categories that are widely dispersed are indicative of a restriction of range effect. Additionally, each rating scale category threshold has an associated outfit mean-square value. The outfit mean-square value is near one when the observed examinee performance measure and expected examinee performance measure of a specific scale category are close. Conversely, the

greater the discrepancy between the observed and expected examinee performance, the bigger the rating scale category's outfit mean-square value will be, which is indicative of restriction of range effect for a rater on that element.

Finally, for the raters still suspected of exhibiting restriction of range effects, the frequency counts were examined to confirm how the rater assigned scores. Moreover, frequency analysis illuminates the nature of restriction of range effects. These analyses provided insight as to whether scores were restricted to the lower or upper ends of the scoring levels, indicating extreme scoring; or the middle scoring levels, indicating a central tendency effect. Raters that showed evidence of a restriction of range effect based on extreme rater infit and outfit values, the spread of rating scale category thresholds and their corresponding extreme outfit values, and/or frequency analyses were determined to be exhibiting restriction of range effect. A total count of such raters was recorded.

Research question 4: Overall, how many raters do not exhibit leniency/severity, halo effect, or restriction of range rater effects? Results from Model 1 (equation 3), Model 2 (equation 4), and Model 3 (equation 5) were used to evaluate this research question. Raters who were not flagged as exhibiting rater effects based on the counts recorded for research questions 1a, 2a, and 3a, were determined to not exhibit leniency/severity, halo effect, or restriction of range rater effects. For the purpose of the study, these raters were identified as preferable candidates for selection of future rating tasks.

Chapter 4: Results

Eight research questions were addressed in this study:

1) Among this group of raters, is there at least one rater exhibiting statistically significant differences in leniency/severity?
    a. If so, which raters exhibit leniency/severity effects?
2) Is there a group-level rater halo effect suggested by the absence of significant differences in the element difficulties?
    a. If so, which raters exhibit halo effects?
3) Is a group-level restriction of range indicated by the absence of significant differences in examinee abilities?
    a. If so, which raters exhibit restriction of range effects?
4) Overall, how many raters do not exhibit leniency/severity, halo effect, or restriction of range rater effects?

For each research question, an MFRM analysis was conducted, separately, on Critical Thinking VALUE Rubric scores and Written Communication VALUE Rubric scores. First, assumption testing was performed on the three formal assumptions of Rasch models: local independence, unidimensionality, and correct model form. Assumption testing was conducted using Model 1.

Assumption Testing

Local independence. Adjusted Yen's Q3 were evaluated to examine the assumption of local independence for residual correlations not exceeding the .20 critical value. Table 4 and Table 5 display the adjusted Yen's Q3 values among Critical Thinking VALUE Rubric elements and Written Communication VALUE Rubric elements,

respectively. No adjusted Yen's Q3 values exceeded the .20 cutoff — satisfying the assumption of local independence for both rubrics.

Unidimensionality. A principal components analysis was conducted on the standardized residuals of each rubric, separately, to evaluate the assumption of unidimensionality for eigenvalues less than 2.0 for each secondary contrast. Table 6 displays the eigenvalues loading on secondary contrasts for the Critical Thinking VALUE Rubric and the Written Communication VALUE Rubric, separately. Eigenvalue loadings on secondary contrasts were less than 2.0, satisfying the assumption of unidimensionality.

Correct model form. The assumption of correct model from was evaluated based on overall model and rater fit. Overall model fit was evaluated based on the absolute value of the standardized residuals where residuals greater than |2.0| indicated highly unexpected scores. Data were determined to fit the specified model well, overall, if less than 5% of the standardized residuals exceeded or were equal to |2.0|. Less than 5% of the standardized residuals exceed |2.0| for both rubric (3.97% and 4.30% of Critical Thinking and Written Communication standardized residuals, respectively), satisfying the assumption of unidimensionality according to overall model fit metrics.

Rater fit was evaluated based on infit (weighted mean squares) and outfit (unweighted mean squares). Rater infit and outfit values within the acceptable range, between 0.5 and 1.5, indicated data fit the specified model well. Values greater than 2.0 were flagged as indicators of major rater misfit. Table 7 displays the rater infit and outfit values that exceed the acceptable range for the Critical Thinking VALUE Rubric and the Written Communication VALUE Rubric, separately. Rater 87 recorded maximum rater

infit and out — the only rater across both datasets exceeding the 2.0 threshold for misfit, which may be because Rater 87 only scored one case. Two Critical Thinking VALUE Rubric raters (out of 118) exceeded the acceptable range for either infit or outfit. Six Written Communication VALUE Rubric raters (out of 104) exceeded the acceptable range for either infit or outfit. These raters provided scores that were less similar to the model-implied scores than predicted by MFRM Model 1. However, because rater fit is also an indicator for the presence of rater effects on the individual-level, these raters were retained in the analysis.

With the assumptions satisfied, the results of MFRM analysis can be presented with confidence.

Evaluation of Research Questions

Research question 1: Among this group of raters, is there at least one rater exhibiting statistically significant differences in leniency/severity? Model 1, specifically the rater facet (α_j), was used to evaluate this research question. 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 for both, the Critical Thinking VALUE Rubric (x^2(117) = 11694.4, p < .01) and the Written Communication VALUE Rubric (x^2(103) = 21192.2, p < .01), suggesting at least one rater differed significantly in leniency/severity from the other raters in the respective groups of raters. For the Critical Thinking sample of raters, the rater separation index (H_o = 5.62) suggested about six statistically distinct levels of rater leniency/severity. For the Written Communication sample of raters, the rater separation index (H_o = 10.01) suggested ten statistically

distinct levels of rater leniency/severity. Moreover, the rater reliability of separation ($R_o = 0.94$ and $R_o = 0.98$, respectively), suggested near-perfect separation and rank-ordering of raters' leniency/severity along the logit continuum.

Research question 1a: Which raters exhibit leniency/severity effects? Model 1, specifically the rater facet (α_j), was used to evaluate this research question. Raters were screened for exhibiting leniency/severity effect based on visual inspection of the Wright map rater facet, as displayed in Figure 2 and Figure 3 for the Critical Thinking VALUE Rubric and Written Communication VALUE Rubric, respectively. Nine raters were flagged for severity (raters 30, 33, 43, 84, and 87 for the Critical Thinking VALUE Rubric and raters 26, 27, 32, and 102 for the Written Communication VALUE Rubric) for deviating in the positive direction from a logit value of 0.0, which represents average rater leniency/severity, more than the other raters. Conversely, eight raters were flagged for leniency (raters 10, 37, and 46 for the Critical Thinking VALUE Rubric and raters 6, 7, 16, 28, and 33 for the Written Communication VALUE Rubric) for deviating in the negative direction from a logit value of 0.0 more than the other raters.

For these raters, severity measures and fair averages were examined, specifically for the deviation between rater's observed average and model expected average, which is displayed in Table 8. Nearly all flagged raters deviated from average leniency/severity by 2 logits. This is contrasted by examples of "normal" raters, not exhibiting leniency severity effects, who have severity measures near zero. These normal raters are discussed and included in tables of flagged raters to demonstrate the contrast between MFRM indicators and frequency counts of raters exhibiting a particular rater effect from a rater that does not. Comparison raters that were most normal according to initial MFRM

indicators were selected for discussion. For instance, as included in Table 8, comparison raters 114 (of the Critical Thinking VALUE Rubric) and 103 (of the Written Communication VALUE Rubric) had severity logit values of -0.02 and 0.09, respectively. However, counter to what was expected, none of the flagged raters with extreme severity values had fair averages that differed greatly from their observed averages. Rather, fair average examination tended not to distinguish flagged raters from comparison raters for exhibiting leniency/severity effect.

For the Critical Thinking VALUE Rubric, differences between observed and fair averages ranged from -0.86 (from Rater 36) to +0.85 (from Rater 81). Note that neither of these raters were flagged for exhibiting rater leniency/severity according to the Wright Map inspection and subsequent logit values. Specifically, of the flagged raters, only Rater 28 was near the extreme end of the range of differences between observed and fair averages. However, Rater 114 (of the Critical Thinking VALUE Rubric), a comparison rater that was not flagged for leniency/severity effect, had less discrepancy between observed and fair averages than most of the raters flagged for leniency/severity effect. Even still, two raters flagged for leniency/severity effect had less discrepancy between their observed and fair averages than the comparison rater.

Interestingly, the differences between observed and fair averages for the Written Communication VALUE Rubric had a smaller range, from -0.54 (from Rater 10) to +0.30 (from Rater 104). With a smaller range, more of the flagged raters were near the extreme ends of the range of differences between observed and fair averages, specifically raters 33, 84, 37, 87, and Rater 10 (of the Written Communication VALUE Rubric), who served as the lower bound of this range. Moreover, Rater 103 (of the Written

Communication VALUE Rubric), a comparison rater that was not flagged for leniency/severity effect, had a similar degree of discrepancy between observed and fair averages as most of the raters flagged for leniency/severity effect. Essentially, fair average examination tended not to distinguish flagged raters from comparison raters for exhibiting leniency/severity effect.

Finally, the frequency counts of raters flagged for leniency/severity were examined to confirm how the raters assigned scores, as displayed in Table 9. Frequency counts presented patterns as expected for severe raters and for lenient raters, albeit with less clarity of distinction. Raters flagged for severity, having logit values greater than 2, tended to assign scores primarily to the two lowest proficiency ratings. For instance, Rater 26 (of the Written Communication VALUE Rubric), with a logit value of 2.31, assigned 34% of their ratings to lowest proficiency level, 52% of their ratings to proficiency level two, and only 10% and 1% to proficiency levels three and four, respectively. Meanwhile, the comparison rater assigned ratings throughout the rating scale, with the bulk of scores assigned to the central proficiency levels. Rater 103 (of the Written Communication VALUE Rubric), a comparison rater not flagged for severity/leniency effect with a logit value of 0.09, assigned 6% of their ratings to the lowest proficiency level, 39% of their ratings to proficiency level two, 39% of their ratings to proficiency level three, and 16% of their ratings to proficiency level four.

The pattern of frequency counts was less distinct for lenient raters; however, raters flagged for leniency, having logit values less than 2, tended to assign more scores to the highest proficiency level and fewer scores to the lowest proficiency level. For instance, Rater 7 (of the Written Communication VALUE Rubric), with a logit value of

-2.27, assigned 30% of their ratings to the highest proficiency level, only 7% of their ratings to the lowest proficiency level, and 22% and 38% to proficiency levels two and three, respectively. Notice how this distribution of ratings is not quite as distinct from Rater 103, the comparison rater described above, as was the pattern identified for the severe raters. Nonetheless, this pattern was consist throughout the flagged raters and more evident for the Critical Thinking VALUE Rubric than the Written Communication VALUE Rubric.

Frequency counts tended to support the flagged raters as exhibiting leniency/severity based on how they assigned scores. As such, the raters initially flagged for exhibiting rater leniency/severity effect from the Wright Map inspection and then supported by the examination of logit values (to the exclusion of rater 87) are determined to be exhibiting rater leniency/severity effect: 10, 30, 33, 37, 43, 46, and 84 raters for the Critical Thinking VALUE Rubric and raters 6, 7, 16, 26, 27, 28, 32, 33, and 102 for the Written Communication VALUE Rubric. Ultimately this classification is a judgement call based on rater performance across all indicators.

Research question 2: Is there a group-level rater halo effect suggested by the absence of significant differences in the element difficulties? Model 2, specifically the element facet (δ_i), was used to evaluate this research question. The fixed-effect chi-square was evaluated to determine whether there were statistically significant differences in element difficulty, after controlling for measurement error. The fixed-effect chi-square was statistically significant for both, the Critical Thinking VALUE Rubric ($x^2(4) = 4293.1$, $p < .01$) and the Written Communication VALUE Rubric ($x^2(4) = 3927.6$, $p < .01$), suggesting at least one element differed significantly in difficulty from the other

elements on the respective rubric. For the Critical Thinking rubric, the element separation index ($H_o = 37.93$) suggested about 38 statistically distinct levels of element difficulty. For the Written Communication rubric, the element separation index ($H_o = 37.19$) suggested 37 statistically distinct levels of element difficulty. Moreover, the element reliability of separation ($R_o = 1.00$ for both rubrics), suggested perfect separation and rank-ordering of element difficulty along the logit continuum.

Research question 2a: Which raters exhibit halo effects? Model 2, specifically the threshold by element facet (τ_{ik}), was used to evaluate this research question. Raters were initially screened for exhibiting halo effect based on rater infit and outfit values. Because we have evidence that element difficulty varied (see results of Research Question 2), raters infit or outfit values greater than 1.5 will be flagged as exhibiting halo effect — displayed in Table 10. Eight raters were flagged for halo effect (raters 41, 42, 84, and 118 for the Critical Thinking VALUE Rubric and raters 3, 48, 49, and 69 for the Written Communication VALUE Rubric) in this initial screening.

A high rater infit or outfit value indicates that the rater had unexpected scores. This may suggest that the rater was not able to differentiate reliably between conceptually distinct traits, specifically assigning similar scores repeatedly over elements of varied difficulty. However, to verify that this is the reason for high infit and outfit, the number of times these raters assigned the same scores throughout elements was calculated — displayed in Table 11. Examination of the frequency with which raters assigned the same score across at least four elements revealed that the high infit and outfit values of two raters (Rater 118 of the Critical Thinking VALUE Rubric and Rater 69 of the Written Communication VALUE Rubric) were more likely due to the few number of cases these

raters scored ($N = 1$ and $N = 2$, respectively). Of the remaining six suspected raters, only two demonstrated a high frequency of the same score assigned across at least four elements: Rater 84 (of the Critical Thinking VALUE Rubric) assigned the same score across at least four elements for 69% of cases and Rater 3 (of the Written Communication VALUE Rubric) assigned the same score across at least four elements for 93% of cases. Meanwhile, the comparison raters not flagged for halo effects, Raters 86 (of the Critical Thinking VALUE Rubric) and 62 (of the Written Communication VALUE Rubric), assigned the same score across at least four elements for 49% and 34% of cases, respectively.

The frequency counts of flagged raters were examined to confirm how the raters assigned scores, as displayed in Table 12. However, frequency counts presented patterns that did not provide the clear confirmation desired. For instance, Rater 84 (of the Critical Thinking VALUE Rubric) assigned scores primarily to two proficiently levels: 64% to level one and 33% to level two. Meanwhile, the comparison rater and the remaining raters flagged for halo effect of the Critical Thinking VALUE Rubric tended to distribute scores across scores more evenly, with 23% assigned to level one, 35% assigned to level two, 32% assigned to level three, and 6% assigned to level four by the comparison rater. In this example, while the pattern of frequency counts exhibited by Rater 84 seems to confirm the rater as exhibiting halo effect, it can also indicate a severe rater or a restriction of range effect.

The patterns observed for the Written Communication VALUE Rubric is even less clear. The comparison rater, Rater 62, had more scores assigned to just two rating levels than any of the other flagged rater for halo effect: 40% assigned to level two and

41% assigned to level three. Moreover, Rater 3, the only rater of the Written Communication VALUE Rubric still suspected of halo effect based on the previous frequency analysis across elements, dispersed ratings across proficiency levels more than the remaining flagged raters. Evidently, evaluation of the frequency counts alone would not clearly identify which raters exhibited halo effect. However, evaluating how frequently raters assigned similar scores elements provided useful information. As such, of the eight raters initially flagged for halo, only two were judged as exhibiting halo effect upon follow-up procedures: Rater 84 (of the Critical Thinking VALUE Rubric) and Rater 3 (of the Written Communication VALUE Rubric). Ultimately this classification is a judgement call based on rater performance across all indicators.

Research question 3: Is a group-level restriction of range indicated by the absence of significant differences in examinee abilities? Model 3, specifically the examinee facet (θ_n), was used to evaluate this research question. The fixed-effect chi-square was evaluated to determine whether there were statistically significant differences in examinee ability, after controlling for measurement error. The fixed-effect chi-square was statistically significant for both, the Critical Thinking VALUE Rubric ($x^2(5108) = 38542.7$, $p < .01$) and the Written Communication VALUE Rubric ($x^2(4287) = 38498.6$, $p < .01$), suggesting at least one examinee differed significantly in ability from the other examinees in the respective groups of examinees. For the Critical Thinking sample of examinees, the examinee separation index ($H_o = 3.60$) suggested about four statistically distinct levels of examinee ability. For the Written Communication sample of examinees, the examinee separation index ($H_o = 4.26$) suggested four statistically distinct levels of examinee ability. Moreover, the rater reliability of separation ($R_o = 0.86$ and $R_o = 0.90$,

respectively), suggested strong separation and rank-ordering of examinees' ability along the logit continuum. Essentially, these results are expected because examinees consist of undergraduate students of various credit levels and, therefore, should differ in ability.

Research question 3a: Which raters exhibit restriction of range effects?

Model 3, specifically the threshold by rater (τ_{jk}), was used to evaluate this research question. Raters were initially screened for exhibiting restriction of range effect based on rater infit and outfit values. Raters with infit or outfit values less than 0.5 or greater than 1.5 will be flagged as exhibiting halo effect, as displayed in Table 13, alongside a comparison rater for each rubric that had infit and outfit values near 1. Three raters were flagged for restriction of range effect (rater 84 for the Critical Thinking VALUE Rubric and raters 69 and 81 for the Written Communication VALUE Rubric) in this initial screening. Rater infit and outfit values outside the acceptable range (from 0.5 to 1.5) may suggest that the raters assigned a wide range of examine ability into a restricted range of the rating scale than most raters.

As such, the rating scale category thresholds and their corresponding outfit values were examined for raters flagged with potentially exhibiting restriction of range effect, along with the comparison raters, displayed in Table 14 and Figures 4, 5, 6, 7, and 8. Counter to what was expected, the raters flagged for restriction of range tended to have less spread in the threshold values than the comparison raters. For instance, Rater 62 (of the Critical Thinking VALUE Rubric; see Figure 4) and Rater 11 (of the Written Communication VALUE Rubric; see Figure 5) were not flagged for restriction of range effect with infit and outfit values near 1. These raters are used as comparisons to the raters flagged for restriction of range effect. The threshold values for the comparison

raters were much wider than for the raters flagged for restriction of range: from -2.51 to 2.85 for comparison Rater 62 and from -2.79 to 2.85 for comparison Rater 11 but only from -1.56 to 1.83 for flagged Rater 69 (of the Written Communication VALUE Rubric; see Figure 6) — the only flagged rater that did not have a missing threshold due no ratings in the highest proficiency level. Nonetheless, the threshold outfit values tended to be considerably worse for the raters flagged for restriction of range. Threshold outfit values for comparison raters were near 1, but were as high as 1.5, 1.8, and even 2.3 for flagged raters.

Moreover, two of the three raters flagged for restriction of range did not have a threshold value between proficiency levels three and four, likely because they did not assign any ratings to proficiency level four. Thus, while the spread of rating scale thresholds does not indicate these raters as exhibiting restriction of range, the threshold outfit values, and the number of thresholds support the interpretation of extreme infit or outfit values in Model 3 as restriction of range.

A high rater infit or outfit value indicates that the rater had unexpected scores and may suggest that the rater overused extreme proficiency levels. A low infit or outfit value indicates that the rater had muted scores, overfitting to model expectations, and may suggest that the rater overused the central proficiency levels. However, to verify that these are the reason for high infit and outfit, the frequency counts of flagged raters were examined to confirm how the raters assigned scores, as displayed in Table 15. Examination of the frequency counts revealed that the extreme infit and outfit values of two raters, Rater 69 and Rater 81 (of the Written Communication VALUE Rubric; see

Figure 7), were more likely due to the few number of cases these raters scored ($N = 2$ and $N = 3$, respectively).

Frequency counts of the only remaining rater suspected of restriction of range presented a pattern of extreme scoring, confirming Rater 84 (of the Critical Thinking VALUE Rubric; see Figure 8) as exhibiting restriction of range effect. Rater 84 assigned 64% of scores to proficiency level one, 33% to level two, only 4% to level three, and assigned no ratings to proficiency level four. Interestingly, the frequency counts of Rater 11 (of the Written Communication VALUE Rubric), one of the comparison raters not flagged for restriction of range, would appear to indicate a restriction of range effect. Rater 11 assigned 29% of scores to proficiency level one, 51% to level two, 19% to level three, and only 1% to level four. However, since the MFRM Model 3 takes into account rater severity, element difficulty, and student ability, it would be erroneous to mark this rater for restriction of range due to frequency counts alone. Thus, frequency counts can be deceptive and misleading. As such, of the three raters initially flagged for halo, only one was judged as exhibiting halo effect upon follow-up procedures: Rater 84 (of the Critical Thinking VALUE Rubric). Ultimately this classification is a judgement call based on rater performance across all indicators.

Research question 4: Overall, how many raters do not exhibit leniency/severity, halo effect, or restriction of range rater effects? Results from Model 1 (equation 3), Model 2 (equation 4), and Model 3 (equation 5) were used to evaluate this research question. Raters who were not flagged as exhibiting rater effects based on the counts recorded for research questions 1a, 2a, and 3a, were determined as not exhibit leniency/severity, halo effect, or restriction of range rater effects. For the

purpose of the study, these raters were identified as preferable candidates for selection of future rating tasks. Out of a total of 221 raters, 17 exhibited evidence of leniency/severity, halo effect, or restriction of range rater effects — seven for the Critical Thinking VALUE Rubric and ten for the Written Communication VALUE Rubric. The pool of preferable raters based on this sample data and MFRM analysis consists of 204 raters: raters 1-9, 11-29, 31, 32, 34, 35, 38-45, 47-83, and 85-118 of the Critical Thinking VALUE Rubric and raters 1, 2, 4, 5, 8-15, 17-25, 29-31, 34-101, and 103, 104 of the Written Communication VALUE Rubric.

Chapter 5: Discussion

Assessment is an essential aspect of educational systems. Institutions of higher education often employee performance assessments, for two reasons in particular: they tend to have increased fidelity to real-world situations and their ability to tap into higher order skills, such as critical thinking written communication (Kuh et al., 2015; Linn, Baker, & Dunbar, 1991). These are skills that are valued across academic disciplines, in the work place, and as life skills (Arum & Roksa, 2011). However, due to the subjective nature of the scoring process, information gleaned from performance assessment can be problematic due to errors in rater judgement (Stiggins, 1987). As such, scores may represent rater tendencies in addition to, or even in place of, examinee ability — which threatens the validity of score use (Engelhard, 2002; Khattri et al., 1998).

Considering the heavy resource demands of administering performance assessments and the serious consequences of score use for students, educators, and educational institutions, it is important that scores are not a function of the raters (Gronlund, 2003; Linn et al., 1991). To this end, robust scoring guides and numerous rater trainings have been employed to root the scoring process in as much objectivity as possible; and still, rater effects persist (Cronbach, 1990).

Evidently, there is a need for additional quality control methods. Moreover, organizations conducting performances assessment, such as AAC&U's VALUE Institute, must employee qualified raters to evaluate artifacts. Deciding who should be selected and how raters should be selected are pressing questions with considerable implications. Thus, the present study was designed to investigate a quality control method of selecting candidates based on the extent to which rater effects influence their judgements.

Specifically, I examined the presence of rater effects in the scores provided by certified VALUE Institute scorers on the two most prominently used rubrics, Critical Thinking and Written Communication, of the popular VALUE rubrics (AAC&U, 2019). The presence of rater effects was evaluated using Many Facets Rasch Measurement for three of the most common and well researched rater effects: leniency/severity, halo, and restriction of range. Raters were determined as exhibiting a particular rater effect based on a culmination of evidence, i.e. several indicators from the MFRM analysis in a stepwise fashion. However, classifications were ultimately judgement calls.

These findings provide context regarding how VALUE certified raters behave and provide insight regarding the utility of various MFRM metrics as indicators of rater effects on the diagnostic level. A general discussion of research findings regarding VALUE Institute scorers and MRFM utility is presented below. Furthermore, limitations and implications of the results, with directions for future research, are discussed.

General Discussion

VALUE Institute Scorers. Overall, the results of this study were fairly positive for AAC&U's VALUE Institute. Most raters were not flagged for exhibiting rater effects. Moreover, several raters were flagged under limited data and consequently were not classified as having sufficient evidence for rater effects. Essentially, most scorers certified by the VALUE Institute are applying the Critical Thinking and Written Communication VALUE Rubrics in a similar and consistent manner, as expected by the MFRM models. Specifically, scorers are distinguishing between elements of varying difficulty and between examinees of differing ability. This can be an indication that current training procedures are working fairly well and builds on the confidence that can

be placed on VALUE Institute scores. Moreover, these findings add to the validity literature supporting the appropriate use of VALUE Rubrics according to the VALUE Institute approach.

Specifically, only 17 raters out of 221 were diagnostically flagged for exhibiting rater effects, seven for the Critical Thinking VALUE Rubric and ten for the Written Communication VALUE Rubric. Most of these raters were flagged for severity/ leniency effects: all seven raters of the Critical Thinking VALUE Rubric flagged were identified as exhibiting severity/leniency effects and nine out of ten raters of the Written Communication VALUE Rubric flagged were identified as exhibiting severity/leniency effects. While the chi-square test of the rater facet, under Model 1, was statistically significant and the rater separation index and rater reliability of separation values indicated that raters could be separated into differing levels of severity in a reliable manner, the fact that only a few raters were identified as lenient/severe may suggest that differential rater leniency/severity is not a persistent problem across most raters. As such, the VALUE Institute may not have to adjust scores for rater leniency/severity. However, the more raters impacted by a high degree of leniency/severity effect, the more the validity of VALUE Institute certified scores may be threatened due to scores representing a function of examinee ability and rater leniency/severity. Therefore, given the present results it may be worthwhile to provide raters with additional training to curb these effects.

The VALUE Institute may want to identify such raters exhibiting rater effects and hold specialized sessions with them to correct their judgement, at a minimum. Rater training may be designed to catch these problems early on. Alternatively, the VALUE

Institute may consider using the Fair Averages provided by MFRM estimation to statistically adjust for rater effects. However, this would require additional explanation in reports to stakeholders for why scores need to be statistically adjusted. Furthermore, key stakeholders may have more confidence in the observed scores provided directly by raters than scores statistically adjusted — gaining buy-in is an important precursor. While the VALUE Institute works with large samples, which are necessary for MFRM analysis, using MFRM requires a specialized program and expertise on how to interpret results. These are barriers that may reduce the feasibility of using MFRM for the VALUE Institute. The VALUE Institute may be able to use the observed averages to help identify severe or lenient raters without the use of MFRM as there was a fairly strong correlation between observed averages and fair averages for both rubrics (87% for Critical Thinking VALUE Rubric and 71% Written Communication VALUE Rubric). However, evaluating raters for rater effects without MFRM may require more screening of individual raters and may have limited success, as was experienced occasionally when examining frequency counts.

Utility of MFRM metrics for diagnosing rater effects. Overall, the results of this study were mixed regarding the utility of MFRM metrics for diagnosing rater effects. The methods used to identify whether at least one rater differed in leniency/severity and to determine if there is evidence of halo or restriction of range effects on the group level were objective with significance tests and effect size indicators. However, the significance tests are likely to be impacted by the large samples sizes that MFRM requires in order to provide parameter estimates with high precision, thus results are likely to be significant often. Moreover, the effect size indicators were not always clear.

For instance, how does one explain and interpret how five elements were separated into over thirty-seven difficulty strata (H_o = 37.93 and H_o = 37.19) for the Critical Thinking VALUE Rubric and Written Communication VALUE Rubric respectively)?

Nonetheless, the MFRM analysis was able to identify individual raters as exhibiting rater effects even for rater effects that were not evident on the group level, such as the halo effect and restriction of range effect. Furthermore, the initial MFRM screen procedures were able to flag the raters most likely to be exhibiting rater effects, effectively reducing the number of rater that needed to be scrutinized from an overwhelming amount (over 200) to a more manageable number (30 were initially flagged in this study). However, while some have some a stronger body of literature behind them (e.g. acceptable infit and outfit ranges), the flagging metrics and further follow-up procedures themselves require subjective judgement to establish.

For instance, there is no clear way to determine where the cut offs should be placed for raters that are too lenient or too severe. For this study, a logit value near |2.0| was selected after examining the distribution of raters along the leniency/severity facet on the Wright Map, whereas a more stringent cutoff would identify more raters for leniency/severity effects. Yet there is no clear guidance on how to determine this cut off value — significance tests, such as t-tests based on rater pairs as recommended by Myford and Wolfe (2004), do not work well due to the large sample sizes. Additionally, Myford and Wolfe (2004) recommended examining the threshold distributions, where greater spread would indicate restriction of range. However, there is no clarity for how spread out the threshold distributions should be. Future research should examine how

different cut scores and varying degrees of standards using the same metric impact the accuracy of classifying raters as exhibiting rater effects.

In addition to the subjectivity of setting the standards for MFRM metrics, considerable subjectivity in determining raters as exhibiting rater effects was necessary, especially without a thorough understanding rubric properties for the various artifact types or expectations for examinee samples. In fact, numerous metrics and methods provided contradictory and ambiguous results. For instance, the deviation in observed averages of raters and fair averages did not provide useful information. This is evidenced with fairly weak correlations between the leniency/severity logit values and the difference between observed and fair averages for both rubrics ($r = 0.37$ and $r = 0.55$ for the Critical Thinking VALUE Rubric and Written Communication VALUE Rubric, respectively), counter to what was expected.

Examining the spread of threshold distributions proved to be even more ambiguous. The thresholds of raters flagged for restriction of range were actually less spread out than most of the raters that were not flagged, which is exactly opposite of how raters were intended to be evaluated for exhibiting restriction of range effect according to the literature. This could potentially be because the raters exhibiting restriction of range tended not to use the fourth and highest proficiency level. Moreover, it may be worthwhile to consider how different threshold patterns may work with central tendency, which is a form of restriction of range. The spread of the thresholds may depend on where the scores are clustered. If scores are clustered at the center then there may be spread at the ends but if restriction is at the low end then the thresholds may not spread as much, especially when dropping an entire proficiency level. However, this is speculative,

and more research is required to understand how threshold distributions should behave under various conditions of restriction of range and under no rater effect patterns.

Nonetheless, the use of MFRM was indispensable for the evaluation of rater exhibiting rater effects. For one, it reduced the number of raters needing to be thoroughly examined for rater effects from an overwhelming amount to a more manageable pool of suspects, which is valuable for practical reasons, especially in large-scale operations. However, the greater utility of MFRM diagnosis is that the analysis takes into account rater severity, element difficulty, and student ability that all influence scores simultaneously. The significance of this was made evident in several of the frequency analyses, most notably the for rater leniency/severity (research questions 1a) and restriction of range (research question 3a).

While the frequency counts, which can be analyzed without MFRM, usually helped clarify how the specific raters were assigning scores, they occasionally appeared deceptive and misleading. For instance, a rater not flagged for restriction of range that was used as a comparison for raters suspected of restriction of range, appeared to indicate a restriction of range effect according to frequency counts due to a lack of scores assigned to the highest proficiency level. And yet, this could be due to the ability level of examinees that the particular rater had as well as the severity of the rater.

Moreover, the frequency counts presented patterns that did not clearly distinguish between raters not exhibiting rater effects and those that did. For instance, the pattern of frequency counts of lenient raters was not always distinct from select, comparative raters not exhibiting leniency/severity effect. A similar lack of clarity was experienced for research question 2a, where the pattern of frequency counts of a comparative rater

seemed to have more evidence of halo effect than raters that were flagged by MFRM indices for exhibiting halo effect. Evidently, these cases warn of the dangers of solely relying on frequency analysis to classify raters for rater effects; specifically, raters that do not have rater effects biasing their judgements may be misclassified as exhibiting rater effects while raters whose judgement is impacted by rater effects may go under the radar and not be identified. This may be an area for future research examining the misclassification rates when relying on MFRM evaluation as compared to relying solely on frequency counts or other competing techniques.

A final note for discussion, stems from the overlap in raters flagged for halo effect and restriction of range effect. Two of the three raters initially flagged for restriction of range (under Model 3) were also flagged for halo effect (under Model 2) based on extreme infit and outfit. It makes sense that a halo effect may be masked as a restriction of range effect, or vice-versa, since they both can appear as similar scores overused, either within an examinee or across examinees — even rater leniency/severity can appear as such. Thus, while the rater effects may be conceptually distinguished and their causes can be distinguished and treated for differently, their detection in MFRM analysis may be muddled.

Overlapping information may be provided from examination of halo and restriction of range MFRM indicators thus the two rater effects may be harder to disentangle. It may require researchers to examine further the nature of the rubric or interview the flagged raters themselves for why they assigned the scores as they did. While this may be an area for future research, it may prove a benefit to subsume the two rater effects, as defined by MFRM indicators, under one analysis step and disentangle the

two with contextual evaluation. Ultimately, a theme of this research is that the classification of raters for exhibiting rater effects requires careful, subjective judgement of the body of evidence as a whole and benefiting from contextual information.

Limitations

VALUE Institute. A limitation of the present study stems from the assessment context of AAC&U's VALUE Institute. A wide range of artifacts are submitted to the VALUE Institute from a variety of universities and colleges. Very little control is exerted over the types of performance assessments or how they are structured and carried out. Thus, the results of a similar analysis on a sample of more well-defined artifacts, perhaps from a single university or program, would yield different results.

Using MFRM. The use of MFRM for diagnosing rater effects has several limitations due to the demands of the technique. As mentioned, a large sample is required to conduct MFRM. Thus, researchers must have the resources necessary to collect and score a large sample of performances assessment artifacts. Otherwise, MFRM may not be feasible, in addition to other practical demands such as skills with specialized software and knowledge of measurement theory to conduct the analysis and evaluate results.

Another limitation of MFRM is that it is a normative technique. In other words, rater estimates are based on the performance of the sample rather than an objective standard. However, there is a difference in behaving like most of the raters in the sample an providing accurate scores. Thus, even though rater 102 of Critical Thinking VALUE Rubric is classified as a severe rater, the rater may actually be applying scores without errors in his judgement whereas the rest of the sample are extremely lenient. Thus, MFRM gives us estimates of rater leniency/severity, element difficulty, and student

ability relative to the sample, which is why additional contextual information may be beneficial to supplement MFRM to confirm accuracy of results.

Conclusion

Often times, the best methods of classification are the ones that require some degree of subjective judgement. In fact, the overwhelming majority of researchers who have provided "cutoffs" as standards for classification or decision making have regretted doing so because it neglects the complexity involved in these situations. For instance, consider the controversy surrounding structural equation model (SEM) fit indices cutoff values. One example is the changing recommendations concerning the root mean squared error of approximation (RMSEA) fit index (Hooper, Coughlan, & Mullen, 2008). RMSEA values from 0.05 to 0.10 were considered a sign of adequate fit and values over 0.10 represented poor fit. At least those were the recommended cutoffs up until the early nineties. Researchers later suggested that RMSEA values between 0.08 and 0.10 indicated mediocre fit and below 0.08 represented good fit. And yet, more recent recommendations from SEM researchers have called for a cut-off value close to .06 or a stringent upper limit of 0.07. Evidently, cutoff guidelines are useful for practitioners and a necessity for developing understanding of the meaning behind indices; however, contextual factors and expert judgement are also key components of interpreting such indices.

Using MFRM metrics for diagnosing rater effects is complex, requiring specialized knowledge to conduct analysis and interpret output, which is a rather subjective process, requiring judgement calls based on an evaluation of several metrics and the evidence as a whole. These indices have "cutoffs" that are also set in some degree

of subjectivity depending how conservative one desires to be. This study has explored an additional quality control method for reducing the presence of rater error in performance assessment scores using several metrics of MFRM analysis on three of the most common rater effects. Ultimately, the methods described need to be refined further and it would be enlightening to see how a pool of raters selected using this method performs regarding the impact of rater effects on the accuracy of scores.

Table 1

Demographic information of VALUE Institute 2018-2019 academic year sample

	Critical Thinking		Written Communication		Overall	
	Count	Percent	Count	Percent	Count	Percent
Sex						
Female	2854	51	2550	52	5404	52
Male	1799	32	1587	33	3386	32
Missing	939	17	747	15	1686	16
Race/Ethnicity						
American Indian or Alaska Native	21	<1	18	<1	39	<1
Asian	215	4	151	3	366	3
Black or African American	266	5	295	6	561	5
Hispanic or Latino	569	10	455	9	1024	10
Pacific Islander	11	<1	3	<1	14	<1
Two or more races	221	4	165	3	386	4
White	3366	60	3062	63	6428	61
Missing	918	16	735	15	1653	16
Age						
Under 19	8	<1	92	2	100	1
19	583	10	738	15	1321	13
20	691	12	800	16	1491	14
21	600	11	603	12	1203	11
22	875	16	755	15	1630	16
23	725	13	527	11	1252	12
24	465	8	305	6	770	7
Over 24	810	15	474	10	1284	12
Missing	835	15	590	12	1425	14
Pell Eligibility						
Eligible	1476	26	1177	24	2653	25
Not-Eligible	2544	46	2373	49	4917	47
Missing	1572	28	1334	27	2906	28
Sector of Institution						
Public, 4-year	3021	54	2457	50	5478	52
Public, 2-year	970	17	793	16	1763	17
Private, 4-year	1501	27	1540	32	3041	29
Missing	100	2	94	2	194	2

Table 2

Interrater reliability for 2015-2016 scores of the VALUE Institute Collaboratives

	Weighted % of exact agreement range	Weighted Cohen's kappa range	Weighted Brennan-Prediger range	Weighted Gwet's AC2 range
Critical Thinking	88-89	.26-.34	.57-.62	.64-.70
Written Communication	88-94	.27-.39	.56-.77	.60-.84

(AAC&U, 2019, p. 32).

Table 3

Summary of model, facet of interest, rater effect indicators and rationale for each research question

Research Question	Model	Facet	Rater effect Indicators	Rationale
1. Among this group of raters, is there at least one rater exhibiting statistically significant differences in leniency/severity?	1	Rater; α_j	Fixed-effect chi-square ($p < .05$); higher H_o & R_o values	Raters as a group, differ in severity levels and can be reliably separated by their severity logits
1a. If so, which raters exhibit leniency/severity effects?	1	Rater; α_j	Deviation on Wright Map, severity measures, fair averages, & frequency counts	A particular rater deviates in severity from an average severity rating
2. Is there a group-level rater halo effect suggested by the absence of significant differences in the element difficulties?	1	Element; δ_i	Fixed-effect chi-square ($p > .05$); lower H_o & R_o values	Elements do not differ in difficulty levels, meaning that all elements were of similar difficulty such that raters were not able to reliability separate them by their difficulty
2a. If so, which raters exhibit halo effects?	2	Thresholds by element; τ_{ik}	$0.5 \geq$ infit & outfit ≤ 1.5; string of same scores; frequency counts	A particular rater had unexpected scores for a particular element and may have repeatedly assigned similar scores across several elements
3. Is a group-level restriction of range indicated by the absence of significant differences in examinee abilities?	1	Examinee; θ_n	Fixed-effect chi-square ($p > .05$); lower H_o & R_o values	Examinees demonstrated similar abilities, meaning that raters assigned similar scores for examinees such that raters did not reliability separate examinees by their abilities

3a. If so, which raters exhibit restriction of range effects?	3	Thresholds by rater; τ_{jk}	0.5 ≥ infit & outfit ≤ 1.5; rating scale category thresholds & outfit; frequency counts	A particular rater had unexpected scores for the particular element and assigned a wide range of examine ability into a shorter rating scale range than most raters
4. Overall, which raters do not exhibit leniency/severity, halo effect, or restriction of range rater effects?	1-3		Not flagged for a rater effect under research questions 1a, 2a, or 3a.	Raters did not exhibit leniency/severity, halo effect, or restriction of range rater effects according to evaluation with MFRM.

Table 4

Adjusted Yen's Q3 values among Critical Thinking VALUE Rubric elements

	Explanation of issues	Evidence	Influence of context and assumptions	Student's position
Evidence	-0.04			
Influence of context and assumptions	-0.14	0.14		
Student's position	-0.10	< 0.01	0.10	
Conclusions and related outcomes	0.12	-0.08	-0.07	0.13

Table 5

Adjusted Yen's Q3 values among Written Communication VALUE Rubric elements

	Context of and Purpose for Writing	Content Development	Genre and Disciplinary Conventions	Sources and Evidence
Content Development	0.18			
Genre and Disciplinary Conventions	-0.01	0.03		
Sources and Evidence	-0.04	0.04	0.04	
Control of Syntax and Mechanics	-0.03	-0.06	0.05	-0.15

Table 6

Eigenvalues loading on secondary contrasts

	Critical Thinking[a]		Written Communication[b]	
	Eigenvalue	% of Variance	Eigenvalue	% of Variance
1	1.55	30.97	1.41	28.28
2	1.34	26.77	1.36	27.18
3	1.07	21.41	1.14	22.76
4	1.01	20.21	1.07	21.35
5	0.03	0.64	0.02	0.43

Note. Analysis was conducted on residuals separately for data from each rubric.

[a] $n = 5138$. [b] $n = 4290$.

Table 7

Rater infit and outfit values that exceed the acceptable range

Rater	Infit	Outfit
	Critical Thinking[a]	
41	1.64	1.55
84	1.41	1.90
87[c]	Maximum	Maximum
	Written Communication[b]	
3	1.87	1.83
28	1.42	1.57
48	1.42	1.57
49	1.60	1.60
69	1.55	1.52
83	1.55	1.52

Note. Analysis was conducted using Model 1 separately for data from each rubric.

[a] $n = 5138$. [b] $n = 4290$. [c] Rater 87 provided scores for only one case.

Table 8

Rater severity and fair average measures of the raters flagged for exhibiting rater effect based on Wright Map inspection, along with comparison raters

Rater	Logit	S.E.	Obs. *M*	Fair *M*	Diff.
Critical Thinking[a]					
10	-2.57	0.11	2.40	2.94	-0.54
30	2.74	0.17	1.34	1.17	0.17
33	2.24	0.09	1.49	1.25	0.24
37	-2.17	0.10	2.39	2.78	-0.39
43	2.23	0.15	1.35	1.25	0.10
46	-2.08	0.06	2.51	2.74	-0.23
84	2.59	0.25	1.40	1.19	0.21
87	1.51	1.93	1.00	1.43	-0.43
114[c]	-0.02	0.08	2.13	1.94	0.19
Written Communication[b]					
6	-2.60	0.06	3.05	3.42	-0.37
7	-2.27	0.05	2.94	3.30	-0.36
16	-1.97	0.05	2.85	3.18	-0.33
26	2.31	0.07	1.77	1.45	0.32
27	2.38	0.12	1.81	1.43	0.38
28	-2.86	0.67	2.80	3.51	-0.71
32	2.12	0.09	1.74	1.50	0.24
33	-2.90	0.06	3.17	3.53	-0.36
102	2.80	0.08	1.58	1.31	0.27
103[c]	0.09	0.10	2.63	2.27	0.36

Note. Analysis was conducted using Model 1 separately for data from each rubric. "Obs. *M"* represents the observed average. "Fair *M"* represents the fair average. "Diff." represents the difference between the obeserved average and the fair average. "S.E." represents the standard error of the logit measure.

[a] $n = 5138$. [b] $n = 4290$. [c] = a comparison rater not flagged for leniency/severity effect.

Table 9

Rater frequency counts of the raters flagged for exhibiting rater leniency/severity effect based on Wright Map inspection, along with comparison raters

Rater	Logit	Score Count				Score Percentage				Cases	Ratings Count		Ratings Percentage	
		1	2	3	4	1	2	3	4	Rated	Assigned	Missing	Assigned	Missing
Critical Thinking[a]														
10	-2.57	48	93	77	37	16	32	26	13	59	255	40	86	14
30	2.74	140	49	7	1	35	12	2	0	80	197	203	49	51
33	2.24	348	151	55	4	43	19	7	0	163	558	257	68	32
37	-2.17	61	102	73	49	20	34	24	16	60	285	15	95	5
43	2.23	160	59	8	2	53	20	3	1	60	229	71	76	24
46	-2.08	130	225	180	150	17	29	23	19	157	685	100	87	13
84	2.59	51	26	3	0	64	33	4	0	16	80	0	100	0
87[d]	1.51	3	0	0	0	60	0	0	0	1	3	2	60	40
114[c]	-0.02	150	454	250	4	17	52	29	0	174	858	12	99	1
Written Communication[b]														
6	-2.60	35	205	348	319	4	22	37	34	189	907	38	96	4
7	-2.27	63	213	364	286	7	22	38	30	190	926	24	97	3
16	-1.97	19	254	487	155	2	27	51	16	190	915	35	96	4
26	2.31	241	368	73	6	34	52	10	1	141	688	17	98	2
27	2.38	109	84	32	16	39	30	11	6	56	241	39	86	14
28	-2.86	0	1	4	0	0	20	80	0	1	5	0	100	0
32	2.12	178	142	56	10	43	35	14	2	82	386	24	94	6
33	-2.90	16	166	330	343	2	19	39	40	171	855	0	100	0
102	2.80	260	213	33	7	49	40	6	1	106	513	17	97	3
103[c]	0.09	15	98	87	39	6	39	35	16	50	239	11	96	4

Note. Analysis was conducted using Model 1 separately for data from each rubric.

[a] $n = 5138$. [b] $n = 4290$. [c] = a comparison rater not flagged for leniency/severity effect. [d] Rater 87 only assigned three scores to one case, meaning that the model is generating estimates on very little information, thus Rater 87 was removed from consideration.

Table 10

Raters flagged for exhibiting halo effect based on extreme rater infit or outfit values, along with comparison raters

Rater	Infit	Outfit
Critical Thinking[a]		
41	1.64	1.55
42	1.52	1.39
84	1.40	1.99
86[c]	1.00	0.98
118	1.53	1.91
Written Communication[b]		
3	1.56	1.55
48	1.62	1.61
49	1.43	1.59
62[c]	1.00	1.00
69	1.82	2.00

Note. Analysis was conducted using Model 2 separately for data from each rubric.

[a] $n = 5138$. [b] $n = 4290$. [c] = a comparison rater not flagged for halo effect.

Table 11

Frequency of same scores assigned across rubric elements of the raters flagged for exhibiting halo effect based on extreme rater infit or outfit values, along with comparison raters

Rater	Count Rated	Same score across 4 elements		Same score across 5 elements		Same score across at least 4 elements	
		Count	Percentage	Count	Percentage	Count	Percentage
				Critical Thinking[a]			
41	89	15	17	0	0	15	17
42	99	29	29	7	7	36	36
84	16	6	38	5	31	11	69
86[c]	81	31	38	9	11	40	49
118	1	0	0	0	0	0	0
				Written Communication[b]			
3	28	23	82	3	13	26	93
48	63	19	30	1	5	20	32
49	68	12	18	4	33	16	24
62[c]	101	29	29	5	5	34	34
69	2	0	00	0	0	0	0

Note. Analysis was conducted using Model 2 separately for data from each rubric.

[a] $n = 5138$. [b] $n = 4290$. [c] = a comparison rater not flagged for halo effect.

Table 12

Rater frequency counts of the raters flagged for exhibiting halo effect based on extreme rater infit or outfit values, along with comparison raters

Rater	Score Count				Score Percentage				Cases	Ratings Count		Ratings Percentage	
	1	2	3	4	1	2	3	4	Rated	Assigned	Missing	Assigned	Missing
Critical Thinking[a]													
41	154	114	111	27	35	26	25	6	89	406	39	91	9
42	222	119	74	21	45	24	15	4	99	436	59	88	12
84	51	26	3	0	64	33	4	0	16	80	0	100	0
86[c]	91	140	131	24	23	35	32	6	81	386	19	95	5
118	3	1	0	0	60	20	0	0	1	4	1	80	20
Written Communication[b]													
3	118	235	164	107	18	37	26	17	128	624	16	98	3
48	27	59	95	126	9	19	30	40	63	307	8	98	3
49	6	94	102	119	2	28	30	35	68	321	19	94	6
62[c]	50	201	206	35	10	40	41	7	101	492	13	97	3
69	1	3	4	1	10	30	40	10	2	9	1	90	10

Note. Analysis was conducted using Model 2 separately for data from each rubric.

[a] $n = 5138$. [b] $n = 4290$. [c] = a comparison rater not flagged for halo effect.

Table 13

Raters flagged for exhibiting restriction of range effect based on extreme rater infit or outfit values and a comparison rater, along with comparison raters

Rater	Infit	Outfit
Critical Thinking[a]		
62[c]	1.01	1.01
84	1.42	1.63
Written Communication[b]		
11[c]	1.00	1.00
69	1.58	1.67
81	1.44	1.86

Note. Analysis was conducted using Model 3 separately for data from each rubric.

[a] $n = 5138$. [b] $n = 4290$. [c] = a comparison rater not flagged for restriction of range effect.

Table 14

Proficiency level thresholds and corresponding outfit values of the raters flagged for exhibiting restriction of range effect based on extreme rater infit or outfit values, along with comparison raters

Rater	One to Two		Two to Three		Three to Four	
	Threshold	Outfit	Threshold	Outfit	Threshold	Outfit
			Critical Thinking[a]			
62[c]	-2.51	1.10	-0.34	1.00	2.85	1.00
84	-1.86	1.80	1.86	0.70	--	--
			Written Communication[b]			
11[c]	-2.79	1.00	-0.05	1.00	2.84	0.80
69	-1.56	0.90	-0.27	1.40	1.83	1.00
81	-1.49	2.30	0.73	1.50	--	--

Note. Analysis was conducted using Model 3 separately for data from each rubric.

[a] $n = 5138$. [b] $n = 4290$. [c] = a comparison rater not flagged for restriction of range effect.

Table 15

Rater frequency counts of the raters flagged for exhibiting restriction of range effect based on extreme rater infit or outfit values, along with comparison raters

Rater	Score Count				Score Percentage				Cases	Ratings Count		Ratings Percentage	
	1	2	3	4	1	2	3	4	Rated	Assigned	Missing	Assigned	Missing
Critical Thinking[a]													
62[c]	121	392	484	121	10	35	44	10	222	1108	2	100	0
84	51	26	3	0	64	33	4	0	16	80	0	100	0
Written Communication[b]													
11[c]	135	227	83	6	29	51	19	1	93	451	14	97	3
69	1	3	4	1	10	30	40	10	2	9	1	90	10
81	1	5	9	0	7	33	60	0	3	15	0	100	0

Note. Analysis was conducted using Model 3 separately for data from each rubric.

[a] n = 5138. [b] n = 4290. [c] = a comparison rater not flagged for restriction of range effect.

Figure 1

Typical rubric features as seen in part of AAC&U's Critical Thinking VALUE rubric

Elements/Dimensions

Proficiency/Score Levels

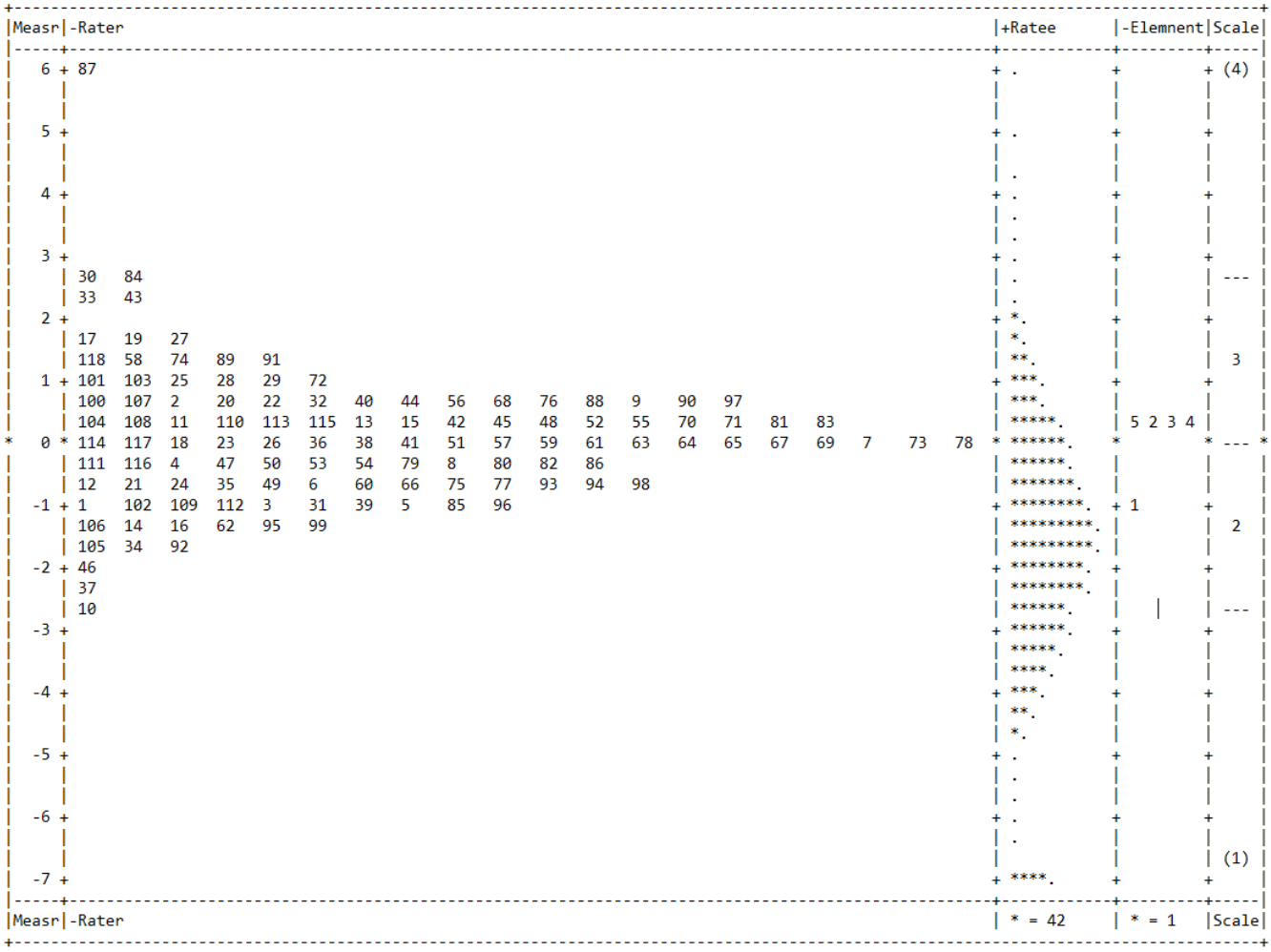
	Capstone 4	Milestones 3	Milestones 2	Benchmark 1
Explanation of issues	Issue/problem to be considered critically is stated clearly and described comprehensively, delivering all relevant information necessary for full understanding.	Issue/problem to be considered critically is stated, described, and clarified so that understanding is not seriously impeded by omissions.	Issue/problem to be considered critically is stated but description leaves some terms undefined, ambiguities unexplored, boundaries undetermined, and/ or backgrounds unknown.	Issue/problem to be considered critically is stated without clarification or description
Evidence	Information is taken from source(s) with enough interpretation/ evaluation to develop a comprehensive analysis or synthesis. Viewpoints of experts are questioned thoroughly.	Information is taken from source(s) with enough interpretation/ evaluation to develop a coherent analysis or synthesis. Viewpoints of experts are subject to questioning.	Information is taken from source(s) with some interpretation/ evaluation, but not enough to develop a coherent analysis or synthesis. Viewpoints of experts are taken as mostly fact, with little questioning.	Information is taken from source(s) without any interpretation/ evaluation. Viewpoints of experts are taken as f act, without question.

Behavioral Descriptor

Scoring Criteria

Figure 2

Rater facet Writght Map of the Critical Thinking VALUE rubric

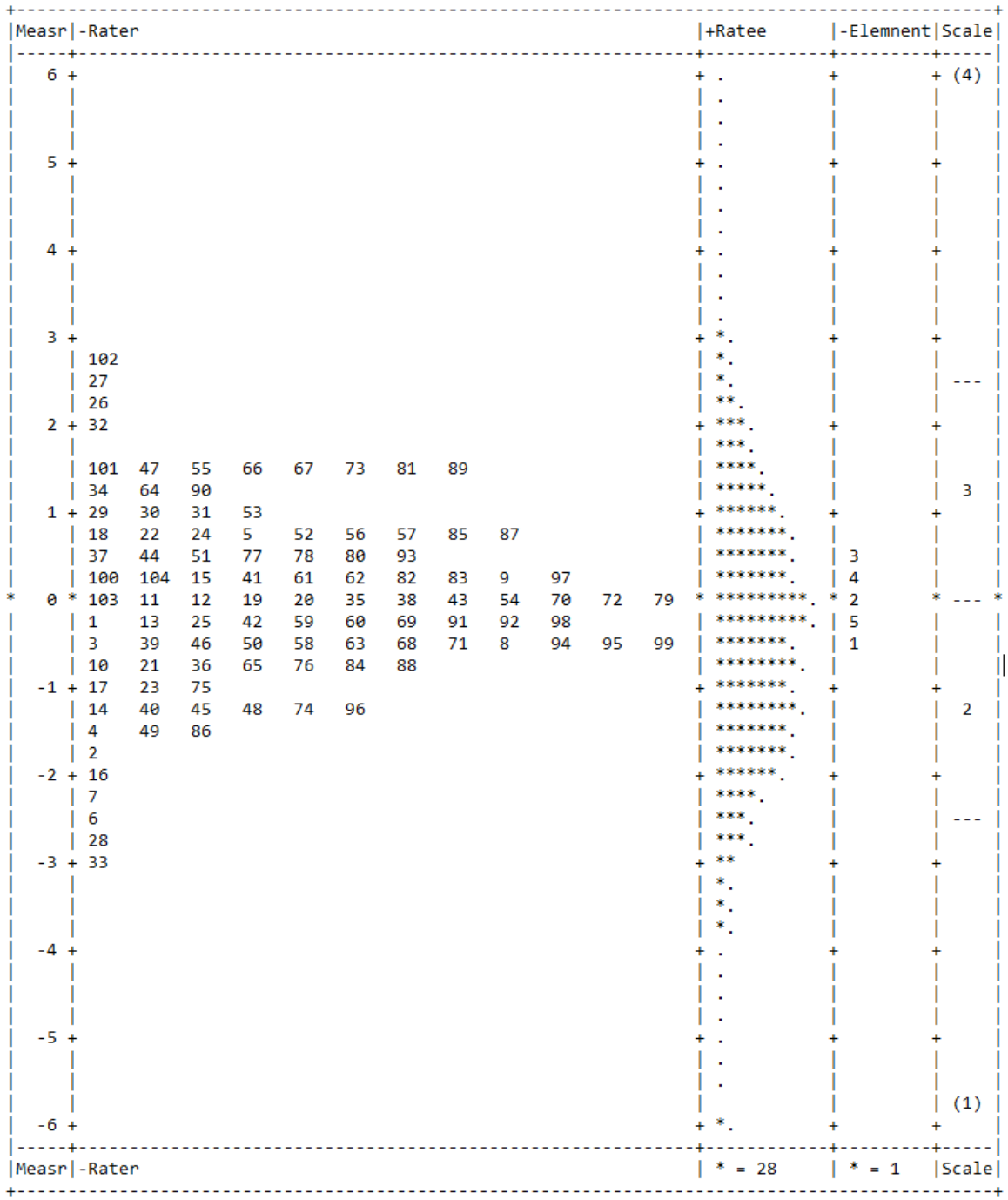


Note. Wright Map generated in FACETS (Linacre, 2017b) output using Model 1. The rater and element facets were centered at 0.00 while the ratee facet was free to vary. The ratee facet was oriented positively, such that higher logit values represent greater ability than lower logit values. The rater and element facets were oriented negatively, such that higher logit values represent more severity and more difficult elements compared to lower logit values, respectively. Rater 87 only assigned three scores to one case, meaning

that the model is generating estimates on very little information, thus Rater 87 was removed from consideration.

Figure 3

Rater facet Writght Map of the Written Communication VALUE rubric

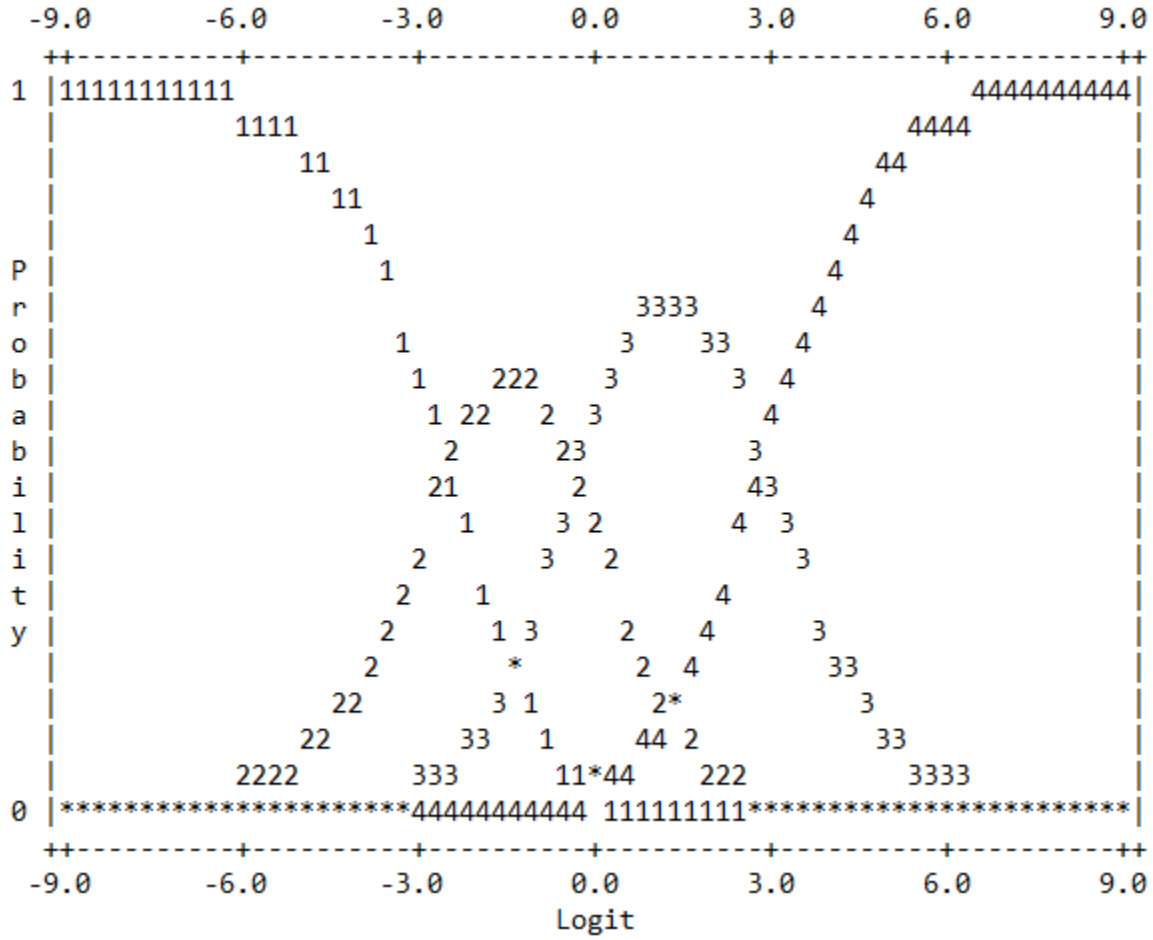


Note. Wright Map generated in FACETS (Linacre, 2017b) output using Model 1. The rater and element facets were centered at 0.00 while the ratee facet was free to vary. The

ratee facet was oriented positively, such that higher logit values represent greater ability than lower logit values. The rater and element facets were oriented negatively, such that higher logit values represent more severity and more difficult elements compared to lower logit values, respectively.

Figure 4

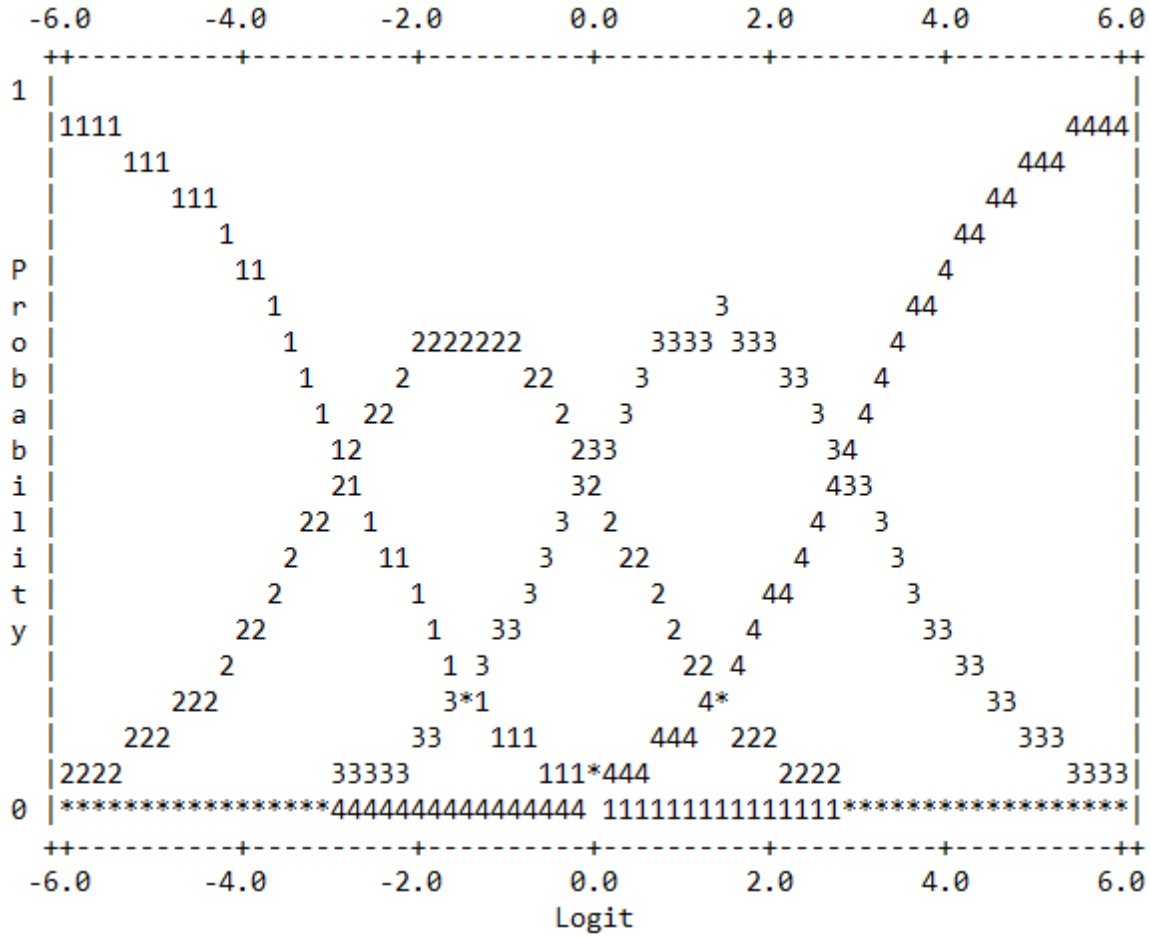
Probability curves of rater 62 of the Critical Thinking VALUE rubric, with infit and outfit values near 1; an example of a rater not flagged for restriction of range effect



Note. Analysis was conducted using Model 3.

Figure 5

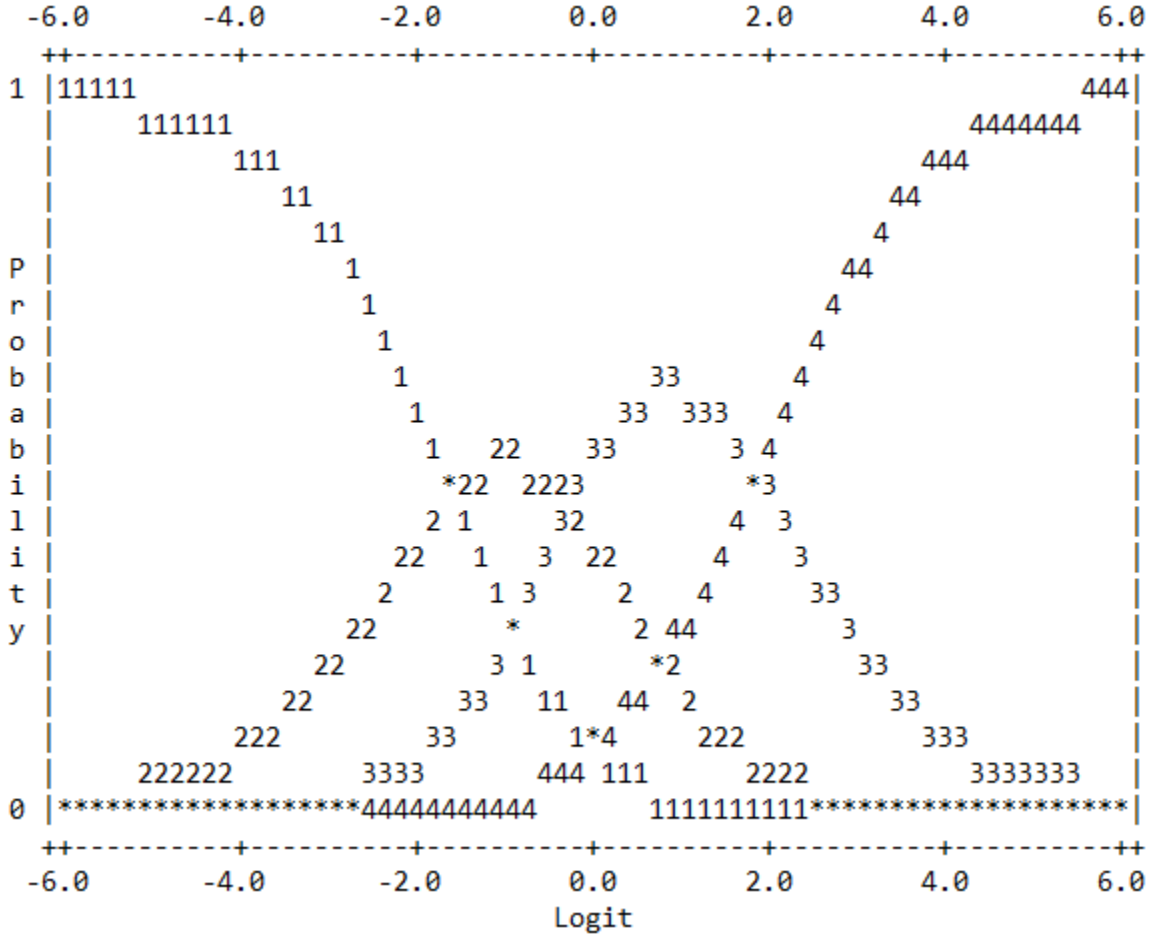
Probability curves of rater 11 of the Written Communication VALUE rubric, with infit and outfit values near 1; an example of a rater not flagged for restriction of range effect



Note. Analysis was conducted using Model 3.

Figure 6

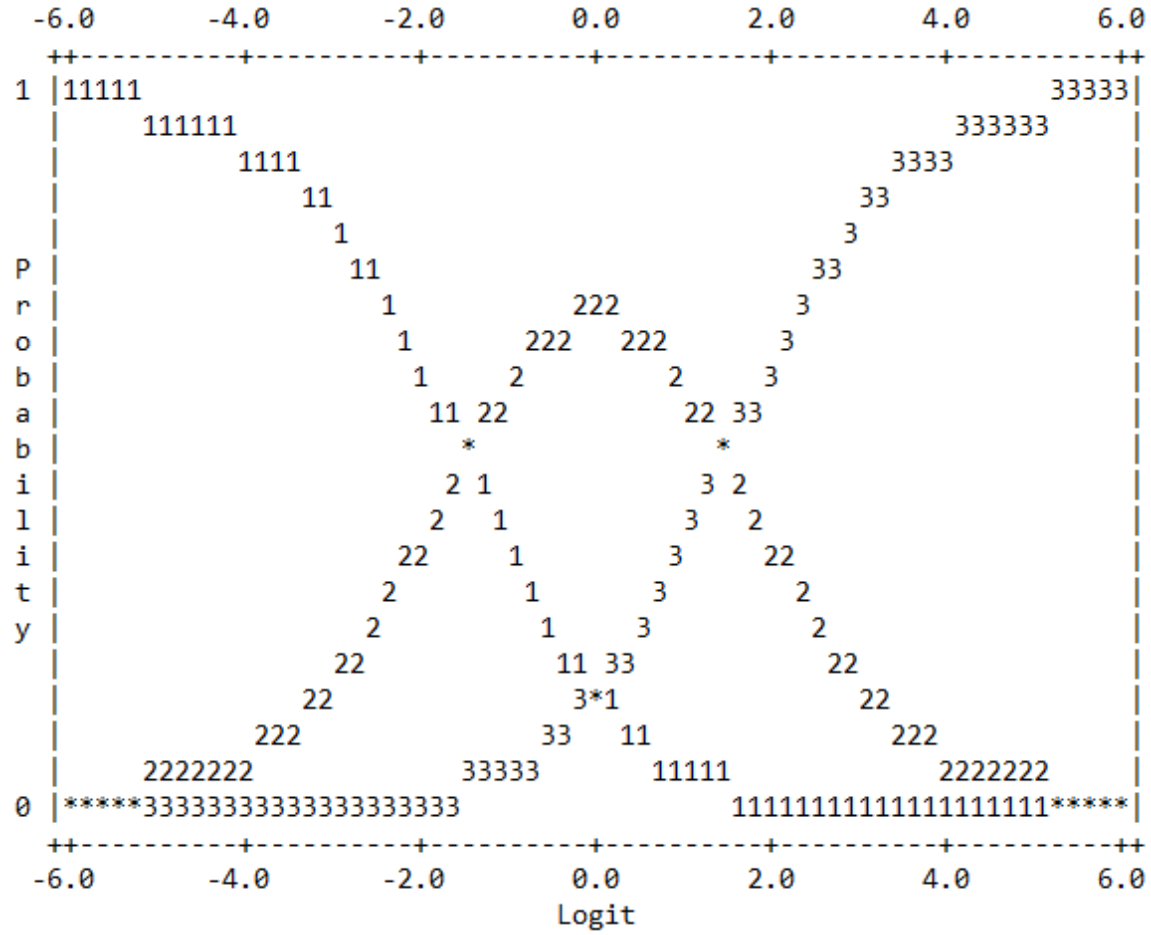
Probability curves of rater 69 of the Written Communication VALUE rubric, flagged for exhibiting restriction of range effect based on extreme infit or outfit values



Note. Analysis was conducted using Model 3.

Figure 7

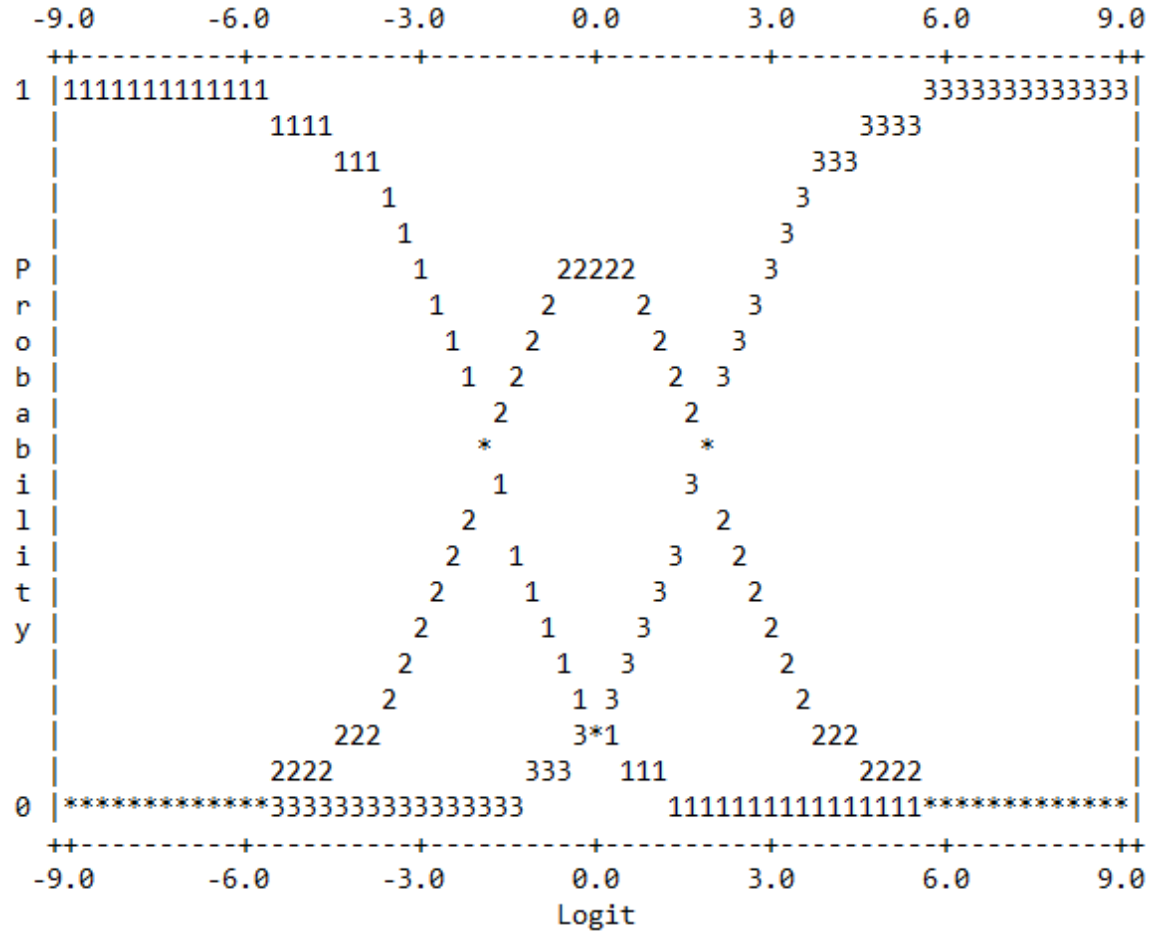
Probability curves of rater 81 of the Written Communication VALUE rubric, flagged for exhibiting restriction of range effect based on extreme infit or outfit values



Note. Analysis was conducted using Model 3.

Figure 8

Probability curves of rater 84 of the Critical Thinking VALUE rubric, flagged for exhibiting restriction of range effect based on extreme infit or outfit values



Note. Analysis was conducted using Model 3.

Appendix A

CRITICAL THINKING VALUE RUBRIC

for more information, please contact value@aacu.org

Association of American Colleges and Universities

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

Critical thinking is a habit of mind characterized by the comprehensive exploration of issues, ideas, artifacts, and events before accepting or formulating an opinion or conclusion.

Framing Language

This rubric is designed to be transdisciplinary, reflecting the recognition that success in all disciplines requires habits of inquiry and analysis that share common attributes. Further, research suggests that successful critical thinkers from all disciplines increasingly need to be able to apply those habits in various and changing situations encountered in all walks of life.

This rubric is designed for use with many different types of assignments and the suggestions here are not an exhaustive list of possibilities. Critical thinking can be demonstrated in assignments that require students to complete analyses of text, data, or issues. Assignments that cut across presentation mode might be especially useful in some fields. If insight into the process components of critical thinking (e.g., how information sources were evaluated regardless of whether they were included in the product) is important, assignments focused on student reflection might be especially illuminating.

Glossary

The definitions that follow were developed to clarify terms and concepts used in this rubric only.

- Ambiguity: Information that may be interpreted in more than one way.
- Assumptions: Ideas, conditions, or beliefs (often implicit or unstated) that are "taken for granted or accepted as true without proof." (quoted from www.dictionary.reference.com/browse/assumptions)
- Context: The historical, ethical. political, cultural, environmental, or circumstantial settings or conditions that influence and complicate the consideration of any issues, ideas, artifacts, and events.
- Literal meaning: Interpretation of information exactly as stated. For example, "she was green with envy" would be interpreted to mean that her skin was green.
- Metaphor: Information that is (intended to be) interpreted in a non-literal way. For example, "she was green with envy" is intended to convey an intensity of emotion, not a skin color.

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone 4	**Milestones** 3	 2	**Benchmark** 1
Explanation of issues	Issue/problem to be considered critically is stated clearly and described comprehensively, delivering all relevant information necessary for full understanding.	Issue/problem to be considered critically is stated, described, and clarified so that understanding is not seriously impeded by omissions.	Issue/problem to be considered critically is stated but description leaves some terms undefined, ambiguities unexplored, boundaries undetermined, and/or backgrounds unknown.	Issue/problem to be considered critically is stated without clarification or description.
Evidence *Selecting and using information to investigate a point of view or conclusion*	Information is taken from source(s) with enough interpretation/evaluation to develop a comprehensive analysis or synthesis.	Information is taken from source(s) with enough interpretation/evaluation to develop a coherent analysis or synthesis.	Information is taken from source(s) with some interpretation/evaluation, but not enough to develop a coherent analysis or synthesis.	Information is taken from source(s) without any interpretation/evaluation. Viewpoints of experts are taken as fact, without question.

	Viewpoints of experts are questioned thoroughly.	Viewpoints of experts are subject to questioning.	Viewpoints of experts are taken as mostly fact, with little questioning.	
Influence of context and assumptions	Thoroughly (systematically and methodically) analyzes own and others' assumptions and carefully evaluates the relevance of contexts when presenting a position.	Identifies own and others' assumptions and several relevant contexts when presenting a position.	Questions some assumptions. Identifies several relevant contexts when presenting a position. May be more aware of others' assumptions than one's own (or vice versa).	Shows an emerging awareness of present assumptions (sometimes labels assertions as assumptions). Begins to identify some contexts when presenting a position.
Student's position (perspective, thesis/hypothesis)	Specific position (perspective, thesis/hypothesis) is imaginative, taking into account the complexities of an issue. Limits of position (perspective, thesis/hypothesis) are acknowledged. Others' points of view are synthesized within position (perspective, thesis/hypothesis).	Specific position (perspective, thesis/hypothesis) takes into account the complexities of an issue. Others' points of view are acknowledged within position (perspective, thesis/hypothesis).	Specific position (perspective, thesis/hypothesis) acknowledges different sides of an issue.	Specific position (perspective, thesis/hypothesis) is stated, but is simplistic and obvious.
Conclusions and related outcomes (implications and consequences)	Conclusions and related outcomes (consequences and implications) are logical and reflect student's informed evaluation and ability to	Conclusion is logically tied to a range of information, including opposing viewpoints; related outcomes (consequences and	Conclusion is logically tied to information (because information is chosen to fit the desired conclusion); some related outcomes (consequences	Conclusion is inconsistently tied to some of the information discussed; related outcomes (consequences

	place evidence and perspectives discussed in priority order.	implications) are identified clearly.	and implications) are identified clearly.	and implications) are oversimplified.

Appendix B

WRITTEN COMMUNICATION VALUE RUBRIC

for more information, please contact value@aacu.org

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

Written communication is the development and expression of ideas in writing. Written communication involves learning to work in many genres and styles. It can involve working with many different writing technologies, and mixing texts, data, and images. Written communication abilities develop through iterative experiences across the curriculum.

Framing Language

This writing rubric is designed for use in a wide variety of educational institutions. The most clear finding to emerge from decades of research on writing assessment is that the best writing assessments are locally determined and sensitive to local context and mission. Users of this rubric should, in the end, consider making adaptations and additions that clearly link the language of the rubric to individual campus contexts.

This rubric focuses assessment on how specific written work samples or collectios of work respond to specific contexts. The central question guiding the rubric is "How well does writing respond to the needs of audience(s) for the work?" In focusing on this question the rubric does not attend to other aspects of writing that are equally important: issues of writing process, writing strategies, writers' fluency with different modes of textual production or publication, or writer's growing engagement with writing and disciplinarity through the process of writing.

Evaluators using this rubric must have information about the assignments or purposes for writing guiding writers' work. Also recommended is including reflective work samples of collections of work that address such questions as: What decisions

did the writer make about audience, purpose, and genre as s/he compiled the work in the portfolio? How are those choices evident in the writing -- in the content, organization and structure, reasoning, evidence, mechanical and surface conventions, and citational systems used in the writing? This will enable evaluators to have a clear sense of how writers understand the assignments and take it into consideration as they evaluate

The first section of this rubric addresses the context and purpose for writing. A work sample or collections of work can convey the context and purpose for the writing tasks it showcases by including the writing assignments associated with work samples. But writers may also convey the context and purpose for their writing within the texts. It is important for faculty and institutions to include directions for students about how they should represent their writing contexts and purposes.

Faculty interested in the research on writing assessment that has guided our work here can consult the National Council of Teachers of English/Council of Writing Program Administrators' White Paper on Writing Assessment (2008; www.wpacouncil.org/whitepaper) and the Conference on College Composition and Communication's Writing Assessment: A Position Statement (2008; www.ncte.org/cccc/resources/positions/123784.htm)

Glossary

The definitions that follow were developed to clarify terms and concepts used in this rubric only.

- Content Development: The ways in which the text explores and represents its topic in relation to its audience and purpose.
- Context of and purpose for writing: The context of writing is the situation surrounding a text: who is reading it? who is writing it? Under what circumstances will the text be shared or circulated? What social or political factors might affect how the text is composed or interpreted? The purpose for writing is the writer's intended effect on an audience. Writers might want to persuade or inform; they might want to report or summarize information; they might want to work through complexity or confusion; they might want to argue with other writers, or connect with other writers; they might want to convey urgency or amuse; they might write for themselves or for an assignment or to remember.
- Disciplinary conventions: Formal and informal rules that constitute what is seen generally as appropriate within different academic fields, e.g. introductory strategies, use of passive voice or first person point of view, expectations for thesis or hypothesis, expectations for kinds of evidence and support that are appropriate to the task at hand, use of primary and secondary sources to provide evidence and support arguments and to document critical perspectives on the topic. Writers will incorporate sources according to disciplinary and genre conventions, according to the writer's purpose for the text. Through increasingly sophisticated use of sources, writers develop an ability to differentiate between their own ideas and the ideas of others, credit and build upon work already accomplished in the field or issue they are addressing, and provide meaningful examples to readers.
- Evidence: Source material that is used to extend, in purposeful ways, writers' ideas in a text.

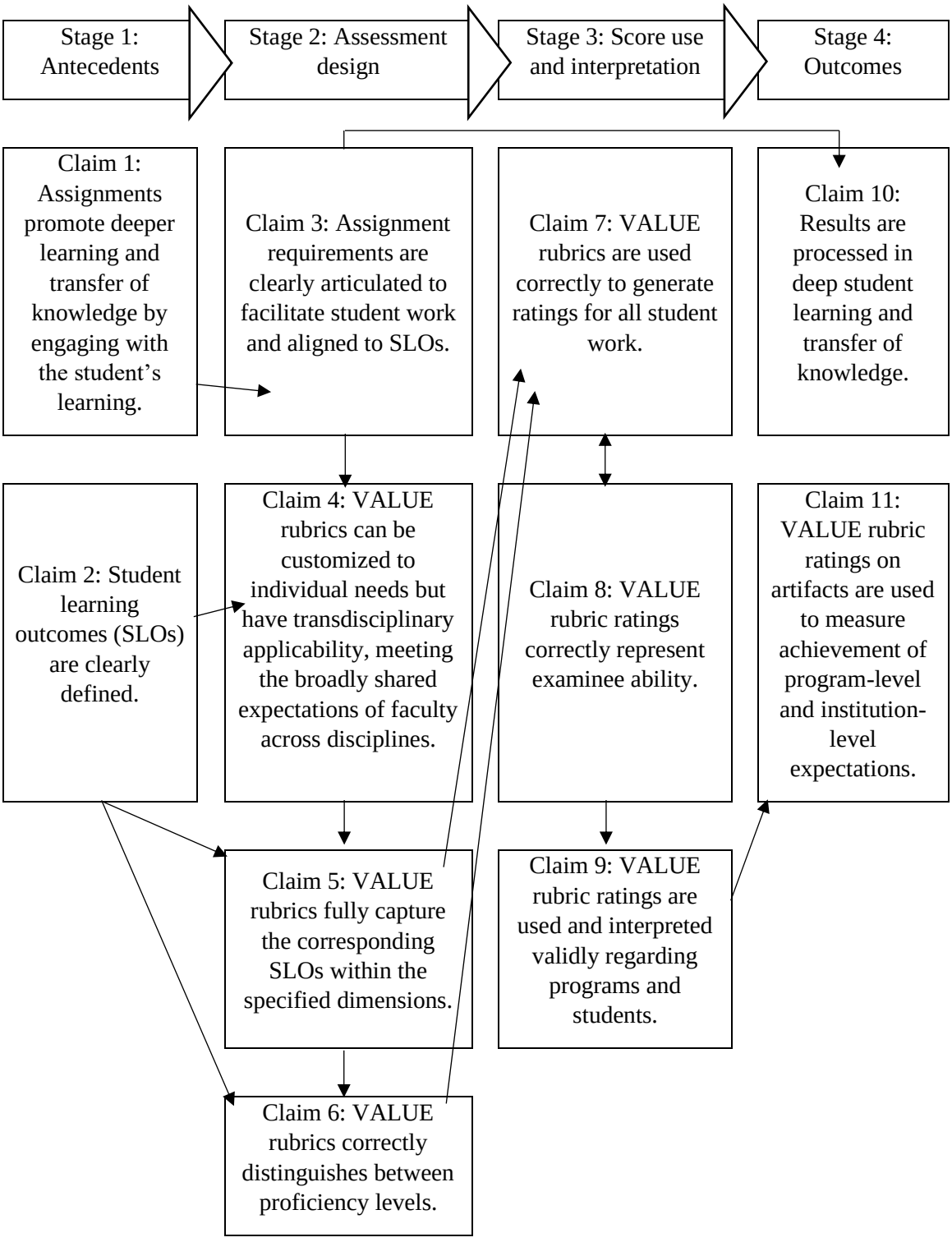
• Genre conventions: Formal and informal rules for particular kinds of texts and/or media that guide formatting, organization, and stylistic choices, e.g. lab reports, academic papers, poetry, webpages, or personal essays.
Sources: Texts (written, oral, behavioral, visual, or other) that writers draw on as they work for a variety of purposes -- to extend, argue with, develop, define, or shape their ideas, for example.

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone 4	**Milestones** 3	2	**Benchmark** 1
Context of and Purpose for Writing *Includes considerations of audience, purpose, and the circumstances surrounding the writing task(s).*	Demonstrates a thorough understanding of context, audience, and purpose that is responsive to the assigned task(s) and focuses all elements of the work.	Demonstrates adequate consideration of context, audience, and purpose and a clear focus on the assigned task(s) (e.g., the task aligns with audience, purpose, and context).	Demonstrates awareness of context, audience, purpose, and to the assigned tasks(s) (e.g., begins to show awareness of audience's perceptions and assumptions).	Demonstrates minimal attention to context, audience, purpose, and to the assigned tasks(s) (e.g., expectation of instructor or self as audience).
Content Development	Uses appropriate, relevant, and compelling content to illustrate mastery of the subject, conveying the writer's understanding, and shaping the whole work.	Uses appropriate, relevant, and compelling content to explore ideas within the context of the discipline and shape the whole work.	Uses appropriate and relevant content to develop and explore ideas through most of the work.	Uses appropriate and relevant content to develop simple ideas in some parts of the work.
Genre and Disciplinary Conventions	Demonstrates detailed attention to and successful execution of a wide range of conventions particular to a specific discipline	Demonstrates consistent use of important conventions particular to a specific	Follows expectations appropriate to a specific discipline and/or writing	Attempts to use a consistent system for basic

Formal and informal rules inherent in the expectations for writing in particular forms and/or academic fields (please see glossary).	and/or writing task (s) including organization, content, presentation, formatting, and stylistic choices	discipline and/or writing task(s), including organization, content, presentation, and stylistic choices	task(s) for basic organization, content, and presentation	organization and presentation.
Sources and Evidence	Demonstrates skillful use of high-quality, credible, relevant sources to develop ideas that are appropriate for the discipline and genre of the writing	Demonstrates consistent use of credible, relevant sources to support ideas that are situated within the discipline and genre of the writing.	Demonstrates an attempt to use credible and/or relevant sources to support ideas that are appropriate for the discipline and genre of the writing.	Demonstrates an attempt to use sources to support ideas in the writing.
Control of Syntax and Mechanics	Uses graceful language that skillfully communicates meaning to readers with clarity and fluency, and is virtually error-free.	Uses straightforward language that generally conveys meaning to readers. The language in the portfolio has few errors.	Uses language that generally conveys meaning to readers with clarity, although writing may include some errors.	Uses language that sometimes impedes meaning because of errors in usage.

Appendix C



Note. Perie's (2013) interpretive argument for VALUE rubrics, adapted from AAC&U (2019, p. 7).

Appendix D

Equations

Equation	Notation	Equation Number
Dichotomous Rasch Model	$ln\frac{P_{ni}}{1-P_{ni}} = \theta_n - \delta_i$	1
Polytomous Rasch Model	$ln\frac{P_{nik}}{P_{nik-1}} = \theta_n - \delta_i - \tau_k$	2
MFRM Model 1	$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_k$	3
MFRM Model 2	$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_{ik}$	4
MFRM Model 3	$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \tau_{jk}$	5
Fixed-effect Chi-square	$x^2 = \sum(W_o * D_o^2) - \frac{(\sum W_o * D_o)^2}{\sum W_o}$	6
True Standard Deviation	$SD_t^2 = SD_o^2 - MSE$	7
Separation Ratio	$G_o = \sqrt{\frac{SD_t^2}{MSE}}$	8
Separation Index	$H_o = \frac{4\sqrt{\frac{SD_t^2}{MSE}} + 1}{3}$	9
Reliability of Separation	$R_o = \frac{\frac{SD_t^2}{MSE}}{1 + \frac{SD_t^2}{MSE}}$	10
Reliability of Separation (2)	$R_o = \frac{SD_t^2}{SD_t^2 + MSE}$	11

Standardized Residual	$Z_{nij} = \frac{x_{nij} - e_{nij}}{\sqrt{w_{nij}}}$	12
Expected Rating	$e_{nij} = \sum_{k=0}^{m} k p_{nijk}$	13
Model Variance	$w_{nij} = \sum_{k=0}^{m} (k - e_{nij})^2 p_{nijk}$	14
Outfit/Unweighted Mean Square	$MS_{U_j} = \frac{\sum_{n=1}^{N} \sum_{j=1}^{I} z_{nij}^2}{NI}$	15
Infit/Weighted Mean Square	$MS_{W_j} = \frac{\sum_{n=1}^{N} \sum_{j=1}^{I} z_{nij}^2 w_{nij}}{\sum_{n=1}^{N} \sum_{j=1}^{I} w_{nij}}$	16

References

American Association of Colleges and Universities. (2019). *We have a rubric for that: The value approach to assessment*. https://www.aacu.org/publications-research/publications/we-have-rubric-value-approach-assessment

Association for American Colleges and Universities. (2017). *On Solid Ground*. Washington, DC: Author.

American Association of Colleges and Universities. (n.d.). *Guide to Develop Your Sampling Plan*.

American Association of Colleges and Universities. (n.d.). *Value Institute Overview*. https://www.aacu.org/VALUEInstitute

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

Archibald, D. A., & Newman, F. M. (1988). Beyond standardized testing: Assessing authentic academic achievement in secondary schools. Washington, DC: National Association of Secondary School Principals.

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

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

Banta, T. W. & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine of Higher Learning, 43*(1), 22-27.

Banta, T. W., Griffin, M., Flateby, T. L., & Kahn, S. (2009, December). *Three promising alternative for assessing college students' knowledge and skills*. (Occasional Paper No. 2). Urbana, Il: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Bartlett, C. J. (1983). What's the difference between valid and invalid halo? Forced-choice measurement without forcing a choice. *Journal of Applied Psychology*, *68*, 218-226.

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

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology, 63*, 301–308.

Bernardin, H. J. & Buckley, M. R. (1981) Strategies in rater training. *Academy of Management Review, 6*, 205-212.

Bernardin, H. J. & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*, 60-66.

Bernardin, H. J. & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62*, 64–69.

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).

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, *64*, 410-421.

Brennan, R. L. (2001). *Statistics for social science and public policy. Generalizability theory*. Springer-Verlag Publishing. https://doi.org/10.1007/978-1-4757-3456-0

Brown, E. M. (1968). Influence of training, method, and relationship on the halo effect. *Journal of Applied Psychology*, *52*, 195–199.

Chickering, A. W. (1999). *Personal qualities and human development in higher education: Assessment in the service of educational goals*. In S. J. Messick (Ed.), *Assessment in higher education: Issues of access, quality, student development, and public policy* (p. 13–33). Lawrence Erlbaum Associates Publishers.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for yen's Q3: Identification of local independence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178-194. 10.1177/0146621616677520

Cobb, G. W. (1998). The objective-format question in statistics: Dead horse, old bath water, or overlooked baby? Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, *90*, 218-244.

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

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.

Dimensionality: contrasts and variances. (n.d.). Retrieved November 11, 2020 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.

Eaton, J. S. (2009). Accreditation in the united states. *New Directions for Higher Education, 145* 79-86.

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. (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 Euro pe/Language Policy Division.

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

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.

Ewell, P. T. (2009, November). *Assessment, accountability, and improvement: Revisiting the tension*. (Occasional Paper No. 1). Urbana, Il: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Gordon, M. E. (1970). The effect of the correctness of the behavior observed on the accuracy of ratings. *Organizational Behavior and Human Performance, 5,* 366–377.

Gough, C. (2020, May 11). *Olympic winter games global tv audience/viewership 2010 to 2018*. Statistia. https://www.statista.com/statistics/531768/global-audience-of-the-winter-olympic-games/

Gregg, Nikole, "Beyond motivation: Differences in score meaning between assessment conditions" (2018). *Masters Theses*. 565. https://commons.lib.jmu.edu/master201019/565

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

Gwet, K. L. (2010). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* (2nd ed.). Gaithersburg, MD: Advanced Analytics LLC.

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

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.

Hart Research Associates (2015). *Falling short? College learning and career success: Selected findings from online surveys of employers and college student conducted on behalf of the Association of American Colleges & Universities*. https://www.aacu.org/sites/default/files/files/LEAP/2015employerstudentsurvey.pdf.

Hathcoat, J. D. (2018). The role of assignments in the multi-state collaborative: Lessons learned from a master chef. *Peer Review*, *20*(4). https://www.aacu.org/peerreview/2018/Fall/Hathcoat

Hedge, J. W., and Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology*, *73*, 68-73.

History.com Editors. (2018, August 21). *The olympic games*. History. https://www.history.com/topics/sports/olympic-games

Holter, L. (2018, February 07). *How do you become an olympic judge? It's a lot of work*. Bustle. https://www.bustle.com/p/how-do-you-become-olympic-judge-its-a-lot-of-work-8147511

Hooper, D., Coughlan, J., & Mullen, M. (2008, June). Evaluating model fit: a synthesis of the structural equation modelling literature. In *7th European Conference on research methodology for business and management studies* (pp. 195-200).

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

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. *Journal of Applied Psychology, 64*, 502-508.

Johnson, R., Penny, J., & Gordon, B. (2009). *Assessing performance: Developing, scoring, and validating performance tasks*. New York: Guilford Publications.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). New York, NY: American Council on Education/Macmillan.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319-342.

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

Landy, F. J., and Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. San Diego, CA: Academic Press.

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.

Latham, G. P., Wexley, K. N., and Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60,* 550-555.

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.

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.

McIntyre, R. M., Smith, D. E., and 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.

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).

Murphy, K. R. (1982). Difficulties in the statistical control of halo. *Journal of Applied Psychology, 67,* 161-164.

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.

Perie, M. (2013). *Developing a validity argument for the VALUE rubrics*. Unpublished manuscript.

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.

Popham, W. J. (1997). What's wrong—and what's right—with rubrics. *Educational Leadership*, 72-75.

Pulakos, E. D., Schmitt, N., and Ostroff, C. (1986). A warning about the use of a standard deviation across dimensions within ratees to measure halo. *Journal of Applied Psychology, 71*, 29-32.

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.

Robbins, S. P. (1989). *Organizational behavior* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

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.

Schmidt, F. L., and Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.

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

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.

Spool, M. D. (1978). Training programs for observers of behaviors: A review. *Personnel Psychology, 31*, 853-888.

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.

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).

Traub, R., & Rowley, G. (1991). NCME Instructional Module: Understanding Reliability. *Educational Measurement: Issues and Practice, 10*, 37-45.

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.

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.

Wexley, K. N., Sanders, R. E., & Yukl, G. A. (1973). Training interviewers to eliminate contrast effects in employment interviews. *Journal of Applied Psychology, 57,* 233–236.

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

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.

Wolf, R. M. (1994). Rating scales. In T. Husen and T. N. Postlethewaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 4923-4930). Oxford, England: Pergamon Press.

Worthen, B. R., White, K. R., Fan, X., and Sudweeks, R. D. (1999). *Measurement and assessment in the schools* (2nd ed.). White Plains, NY: Addison Wesley Longman.

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.