Distribution Agreement

Signature: ____ Date: __________

Stefanie A. Wind

Evaluating the Quality of Ratings in Writing Assessment:

Rater Agreement, Error, and Accuracy

By

Stefanie A. Wind
Master of Arts
Educational Studies, Quantitative Methodology

George Engelhard, Jr., Advisor

Yuk Fai Cheong, Committee Member

Robert J. Jensen, Committee Member

Accepted:

Lisa A. Tedesco, Ph.D. Dean of the James T. Laney School of Graduate Studies

Date

Evaluating the quality of ratings in writing assessment:

Rater agreement, error, and accuracy

By

Stefanie A. Wind
Bachelor of Arts
Bachelor of Music

Advisor: George Engelhard, Jr.

An abstract of
A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University in partial fulfillment of
the requirements for the degree of
Master of Arts
in Educational Studies: Quantitative Methodology
2012

September 12, 2011

Abstract

Evaluating the quality of ratings in writing assessment:

Rater agreement, error, and accuracy

By

Stefanie A. Wind

The purpose of this study is to examine the congruence among methods used to evaluate the quality of ratings obtained in large-scale performance assessments. Within the context of a large-scale writing assessment, this study focuses on the alignment between operationally used indices of rater agreement, error and systematic bias, and direct measures of accuracy within a traditional and Rasch-based approach. This study uses 365 essays from the Georgia High School Writing Test that were rated by 20 operational raters and by a committee of "expert raters," whose scores were used to compute direct accuracy measures. The Facets computer program (Linacre, 2010) is used to compute all of the indices of rating quality. Major empirical findings suggest that Rasch-based indices of model-data fit for ratings as well as indices of rater agreement from Facets (Linacre, 2010) provide information about raters that is comparable to direct measures of accuracy. Because direct measures of rater accuracy are often not attainable in operational settings, the use of easily obtained approximations of direct accuracy measures holds significant implications for monitoring rating quality in large-scale rater-mediated performance assessments.

Evaluating the quality of ratings in writing assessment:
Rater agreement, error, and accuracy

By

Stefanie A. Wind
Bachelor of Arts
Bachelor of Music

Advisor: George Engelhard, Jr.

A thesis submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University in partial fulfillment of
the requirements for the degree of
Master of Arts
in Educational Studies: Quantitative Methodology
2012

September 12, 2011

Table of Contents

List of Tables and Figures

Appendix A

Evaluating the quality of ratings in writing assessment:
Rater agreement, error, and accuracy

The quality of ratings assigned in performance assessments must be systematically examined to ensure high-quality scores for all students. Previous research on rater-mediated assessments has not compared the evaluative criteria of rater agreement, error, and accuracy within traditional and Rasch-based approaches in order to determine the congruence among these indices of rating quality. This study examines relationships among indices of rating quality within a traditional and Rasch-based approach in order to determine the degree to which consistent information is provided about a set of raters regarding the quality of their ratings.

Evaluative criteria for the quality of ratings used within traditional and Rasch-based approaches include measures of rater agreement, errors, and accuracy to make inferences regarding the quality of ratings. When indices result in different conclusions, the quality of the ratings becomes questionable. A need for clarity and consistency in the information provided by evaluative methods for ratings necessitates invariant measurement—the requirement that the same construct be measured across the entire population (in this case, the entire population of raters). In this study, models based on principles of invariant measurement are used along with Brunswik's (1952) lens model for examining perceptions of reality to compare the implications of evaluative criteria for rating quality.

Theoretical Framework

Invariant Measurement

The theoretical framework for this study is based on invariant measurement. Invariance is a principal concept for measurement in the physical and psychological sciences, and the quest for invariant measurement has deep historical roots (Engelhard, 2008). Thurstone, an early 20th century researcher in psychopsychics and psychometrics, recognized the need for objectivity through invariant measures. Describing objective measurement, he wrote, "the scale must transcend the group measured. A measuring instrument must not be seriously affected in its measuring function by the object of measurement…its function must be independent of the object of measurement" (Thurstone, 1928, p. 547).

Invariant measurement is not directly observable; rather, it is a hypothesis that must be confirmed or disconfirmed by evidence in a data set (Engelhard, 1994). Invariance is fundamental to Rasch's idea of *specific objectivity*, which he describes as a requirement for measurement. According to Rasch (1968), if specific objectivity is not achieved,

> the conclusion about, say, any set of person parameters will depend on which other persons are also compared. As a parody we might think of the comparison of the volumes of a glass and a bottle as being influenced by the heights of some of the books on a shelf. Secondly, the conclusions about the persons would depend on just which items were chosen for the comparison, a situation to which a parallel would be that the measurement of the relative height of two persons would depend on whether the measuring stick was calibrated in inches or in centimeters (pp. 29-30).

Wright and Stone (1979) describe the concept of invariant measurement in terms of requirements for the measurement of persons and items. Engelhard and Perkins (2011) expand the conditions for invariant measurement to five requirements related to person measurement, item calibration and dimensionality of measurement. Adherence to these requirements in data can be used as evidence of invariant measurement for persons and items in an assessment situation. The requirements, given in Engelhard and Perkins (2011) are as follows:

1. The calibration of the items must be independent of the particular persons used for calibration: *Person-invariant calibration of test items.*
2. Any person must have a better chance of success on an easy item than on a more difficult item: *Non-crossing item response functions.*
3. The measurement of persons must be independent of the particular items that happen to be used for the measuring: *Item-invariant measurement of persons.*
4. A more able person must always have a better chance of success on an item than a less able person: *Non-crossing person response functions.*
5. Persons and items must be located on a single underlying latent variable: *Unidimensionality* (p. 41).

When raters are introduced into the measurement process, the third requirement is particularly relevant for addressing issues related to the quality of ratings in writing assessment. This requirement can be restated as follows:

- The measurement of persons must be independent of the particular raters that happen to be used for the measuring: *Rater-invariant measurement of persons* (Engelhard, 2002).

Rasch Measurement Theory

Rasch (1960/1980) designed a probabilistic model that meets the requirements for invariant measurement. The Rasch model and its extensions conceptualize measurement in terms of a construct represented as a single line that can be used to explain differences in person achievement and item difficulty. Bond and Fox (2007) describe the Rasch model as a method for determining whether an assessment meets the requirements for invariance, stating, “The Rasch model incorporates a theoretical idealization (or construct, or fiction) of the data’s interrelations, an unachievable state that is mathematically represented as the ideal straight line” (p. 36). When acceptable model-data fit is observed, the requirements for invariance have been met. In other words, a match between observations in data and expectations by the Rasch model indicates that the construct of interest is being measured without interference by construct-irrelevant factors, such as individual rater characteristics, or dependency on construct-irrelevant characteristics of an assessment situation, such as items or persons.

Rather than interpreting observed achievement in terms of raw scores, Rasch measurement theory transforms ordinal observations to modeled latent variable-location estimates for persons and difficulty estimates for items. Based on the differences between item and person locations, items can be judged in terms of their usefulness for providing information about specific persons at varying levels of hypothesized achievement. In addition, the level of precision of the outcomes provided by an instrument can be determined for both persons and items based on the match between the locations of persons and items presented in the graphical display provided in most Rasch measurement software, including Facets (Linacre, 2010), which is the program used in this study. Facets produces a *variable map* in the output for Rasch analyses, which displays location estimates for each facet on a common vertical ruler.

Various models based on Rasch measurement theory can be used to describe the relationship among facets of an assessment situation, including persons and items. The Many-Facet Rasch Model (MFRM) is applied in this study (Linacre, 2010). The MFRM is an extension of Rasch measurement models used to depict the relationship between facets of an assessment situation and the probability (described in terms of the log of the odds, or logit) of observing specific outcomes within assessment situations involving multiple facets, such as individual raters domains in an analytic rubric. In the context of a rater-mediated assessment, the model can be expressed mathematically as:

$$ln\left[\frac{P_{nijx}}{P_{nijx-1}}\right] = \theta_n - \lambda_i - \delta_j - \tau_x, \tag{1}$$

where

P_{nijx} = probability of student *n* receiving a rating of *x* by rater *i* on domain *j,*

P_{nijx-1} = probability of student *n* being rated *x-1* by rater *i* on domain *j,*

θ_n = writing achievement of student *n,*

λ_i= severity of rater *i,*

δ_j= difficulty of domain *j,* and

τ_x= difficulty of rating step *x* relative to step *x-1.*

The MFRM is an additive linear model that provides estimates of achievement on a linear logistic scale (logits) that ranges from positive to negative infinity. The first three variables on the right side of the equation represent facets of the assessment environment. Rasch measurement theory assumes that each person can be characterized by a specific location on the latent variable (construct) and each item by a level of difficulty that can be represented through quantitative values and placed on the same line. The difference in locations between persons and items allows the probability for an observation of a specific score to be estimated. In the context of rater-mediated writing assessment, students are calibrated in terms of writing achievement, raters in terms of severity, and the domains in an analytic rubric in terms of difficulty on the same linear logistic scale. The transformation to the logit scale provides comparable estimates of the relative location of each of these facets to the other students, raters, and domains in the sample, respectively. As can be seen in this equation, the differences in the location of each of these three facets determine the likelihood for a particular student to receive a particular score from a particular rater within a particular domain. The fourth element on the right side of the above equation, τ_x, represents the location of the threshold between categories *x* and *x-1*. In the context of rater-mediated writing assessment, τ_x represents an estimate of the difficulty of movement between adjacent rating scale categories. When good model-data fit is observed, estimates of achievement are considered invariant over the facets of the assessment situation (Engelhard, 1992).

Brunswik's Lens Model

Brunswik (1952) proposed a lens model for probabilistic functionalism to describe perception that has been applied to judgment analysis (Hogarth, 1987; Reynolds & Gifford, 2001). His concept of *probabilistic functionalism* is based on the premise that the probability for an organism to complete an action depends on the nature of their environment. According to his model, aspects of an organism's environment are "lenses," or mediating variables, through which events are perceived. In his words:

> perceptual cues and behavior means are like "signals" in "coded messages." The perceived objects and behavioral results which correspond to the message are mediated through "noisy channels." These latter are contaminated with interferences or constraints of their own which reduce the sender's freedom of choice … in the medium that must be molded to carry out the message…The inherent tangledness of the causal texture of the environment of a behaving organism may be seen as a specific type of "noise"…the undesirable uncertainty arising from structural or statistical properties of the medium is in inverse relationship to the desirable uncertainty which arises by virtue of freedom of choice of the message to be transmitted (Brunswik, 1952, p. 91).

Positioning behavior as a function of context, Brunswik describes behavior as complex and subtle, and calls for consideration of an organism's environment as "a *causal texture* providing the conditions and supports for the organism's behavior" (Postman & Tolman, 1959, p. 508, [italics in original]).

Brunswik's lens model for probabilistic functionalism is depicted graphically in Figure 1. The basic premise of this model is that elements of an organism's environment mediate its behavior, thus potentially distorting the accuracy of judgment about that behavior. The premise underlying the lens model, that reality may be distorted by error, or mediating variables, can be directly related to the goal of invariance in psychological measurement. Rasch (1960/1980) describes his model as a method for identifying relevant and ignoring unnecessary information about persons and items based on imperfect information from an assessment that reflects Brunswik's model. He writes:

> It should not be overlooked that the laws do not at all give an accurate picture of nature. They are simplified descriptions of a very complicated reality. The laws describe an ideal universe, a model, on which reality may be mapped – leaving aside a lot of details. E.g. "a heavy point swinging without friction in a weightless string" never existed in the real world, but at a certain stage of the process of knowledge it is a very useful model of a pendulum. This lesson seems worth keeping in mind when we attempt to build up models for what seems much more complicated, the behavior of human beings (p.10).

Brunswik's lens model can adapted for use in the context of a rater-mediated writing assessment. When the lens model is applied to rater-mediated writing assessments, raters can be considered the lens through which student writing achievement becomes observable. According to Landy and Farr (1980), the idea that context plays a role in judgment necessitates the consideration of unique and group-level cognitive processes in the interpretation of ratings. Ratings, they assert, must be interpreted in light of the fact that “all information must ultimately pass through a cognitive filter represented by the rater,” and the use of multiple raters in performance assessment implies “multiple filters that combine in some particular manner” (p. 100). Considering a rater-mediated writing assessment in terms of Brunswik’s model, mediating cues may interfere in the assessment situation and distort rating quality. These cues may include characteristics of students (e.g., gender, race/ethnicity, best language or opportunity to learn), raters (e.g., error or systematic bias), the assessment (e.g., domain difficulty), or even the evaluative measures used to judge rating quality (e.g., indices of rater agreement, rater errors/biases, or rater accuracy).

Significance of the Study

An individual’s ability to effectively communicate through written text holds significant implications in the 21st century. Because of its growing importance, the assignment of high-quality ratings to written composition is often subject to debate. When assessments hold high-stakes implications, methods in place for evaluating the quality of scores become critical. The idea of a “washback” effect suggests that high-stakes assessments influence the content as well as the pedagogical methods used in instruction (Hamp-Lyons, 1991, 2002; Messick, 1989). According to Hamp-Lyons (2002), the combination of the subjective nature of writing with high-stakes assessments substantiates writing tests as “a form of social engineering that is at once beneficial and dangerous” (p. 5). Assessments that emphasize specific purposes for composition or types of composition, she asserts, are in effect extinguishing the techniques not promoted and, as a result, introducing potentially different implications for student subgroups. Similarly, in an examination of the focus of writing instruction over the last 50 years, Hillocks (2005) notes the persistent power of writing assessment criteria, even when ill-defined, to shape curricular content.

Purpose of the Study

The purpose of this study is to examine and compare indices of the quality of ratings in performance assessments of writing achievement as demonstrated by levels of rater agreement, error and systematic bias, and accuracy. This study applies methods from these three categories of indices to data from a statewide assessment of writing. Considering ratings at the domain-score level, this study seeks to empirically examine and compare indices of rating quality based on rater agreement, error and systematic bias, and accuracy.

Research Questions

This study is guided by two main research questions:

- What are the major indices of rater agreement, error and systematic bias, and accuracy that can be used to evaluate the quality of ratings in rater-mediated assessments?
- Do indices of rating quality within traditional and Rasch-based approaches identify raters in the same way in terms of rating quality?

The first question is addressed through a review of the performance and writing assessment literature. The second question is addressed through the use of empirical data to evaluate operational rating data using selected rating quality indices. Correlational analyses and visual displays are used to examine relationships among indices.

Definitions

Following are definitions of key terms used frequently throughout the study.

Expert Raters and Validity Committees – individual or groups of raters whose expertise is considered sufficient for the assignment of scores that reflect “true” or “accurate” measures of an student’s achievement in terms of specified criterion for a specified construct. Scores assigned by expert raters and validity committees are used as criterion for the evaluation of the quality of scores assigned by operational raters (defined below).

Facets – a software program used to conduct analyses of rater judgments based on Rasch measurement theory. Version 3.67.0 (Linacre, 2010) of Facets is used for analyses in this study.

Operational raters – raters who have completed training for a specific assessment context and evaluate or judge the quality of student performances according to specified criterion.

Performance assessments – assessments that require students to carry out a task rather than select an answer from a list of choices. Examples include essay assessments of writing or the creation of portfolios.

Percent exact rater agreement – a measure of the degree to which raters assign equivalent scores in the same circumstance, i.e., the degree to which ratings are affected by random variation unrelated to the construct of interest. Rater agreement occurs when a population of raters assigns similar scores to students.

Rater accuracy – the degree to which raters assign scores that are equivalent to those assigned by individual or committees of expert raters, considered to reflect the hypothesized “true” score for a given student.

Rater error and systematic bias – random and systematic variation in scores that occur as a result of influences of construct-irrelevant factors on evaluation of a performance. Rater errors and systematic biases are thought to contribute to the assignment of scores different than those warranted by performance.

Rater-mediated assessments – assessments that require human judgment for the assignment of scores to a performance. These assessments typically involve the use of pre-determined criterion against which student performance is judged.

Rater severity – a type of rater error characterized by a rater's tendency to assign scores that are lower than those warranted by student performance.

True Score – a hypothetical score that perfectly relates a student's achievement on a specified construct to a score category.

Review of the Literature

This section presents a review of literature on rater-mediated assessment that concerns three approaches for evaluating the quality of ratings: (a) rater agreement, (b) rater error and systematic bias and (c) rater accuracy. After an overview of research on rater-mediated writing assessment, each approach for evaluating rating quality is discussed in terms of historical and current applications from both a traditional and Rasch measurement theory-based approach.

Rater-Mediated Writing Assessment

Rater-mediated assessments require human judgment for the assignment of scores to performance. Driven by a widespread assumption that performance assessment tasks require authentic application of a construct and provide more valid information about student achievement than do indirect applications (e.g., multiple-choice items), these assessments are widely used to measure writing achievement (Hillocks, 2005). However, the quality of rater-mediated assessments is challenged by the use of human judgment for the assignment of scores to performance tasks. The requirement of human judgment for the assignment of scores has implications for the inferences that can be made from their results. According to Engelhard

(2002), rater-mediated assessments "do not provide direct information regarding student achievement because the student's responses must be mediated and interpreted through raters to obtain judgments about student achievement" (p. 261). Unlike multiple-choice items aimed at assessing skills assumed to reflect writing achievement, essay items are often considered direct measures of writing achievement because students are asked to create a sample of their writing that can be scored by human raters with a rubric (Hillocks, 2005). As a result, essay-based assessments of writing are known as *rater-mediated performance assessments* (Engelhard, 2002). As the term *rater-mediated* implies, it is important to note that, although students are performing the act of writing an essay, their scores are actually *rater interpretations* of writing achievement based on an assessment artifact.

Indices of Rating Quality in Writing Assessment

Methods used to evaluate the quality of ratings in performance assessment are, essentially, attempts to evaluate the degree to which ratings are reflections of "true" achievement. Statistical analyses provide a variety of methods for estimating the presence of model-defined true scores that have been applied to rater-mediated writing assessments. However, conceptual issues challenge reliance on statistical approaches to evaluate rating quality. Describing a conflict between procedures from physical sciences in which stability and independence of behavior are assumed to be possible and the lack of these two properties in psychological science, Cronbach (1947) limits the ability for statistical procedures to adequately capture "truth" and "error" in test scores. He opines:

> 'Chance' and 'error' are merely names we give to our ignorance of what determines an event. All methods of studying reliability make a somewhat fallacious division of variables into 'real variables' and 'error'… A test score is

> made up of all these ‘real’ elements, each of which could be perfectly predicted if our knowledge were adequate. Reliability, according to this conception, becomes a measure of our ignorance of the real factors underlying brief fluctuations of behavior and atypical acts (pp. 6-7).

Despite this conceptual conflict, statistical indices are used in practice to evaluate rating quality by identifying individual or groups of raters who appear to be functioning differently from a group, or individual or groups of students who appear to be consistently rated differently from the population of students in a specific assessment context.

Murphy and Cleveland (1991) describe three broad categories for the evaluation of rating quality in performance assessments: (a) rater agreement, (b) rater error and systematic bias and (c) rater accuracy. Research on the quality of ratings within rater-mediated assessment includes the development and application of a variety of measures within each of these categories. The following discussion reviews methods for rating quality within a traditional approach and one based in Rasch Measurement Theory.

Rater Agreement

In practice, reliability coefficients for measures of rater agreement are used as indices of rating quality based on a traditional approach (Johnson, Penny, & Gordon, 2009). Classical Test Theory defines reliability as an estimation of the proportion of observed score variance attributable to true score variance (Crocker & Algina, 2008). In his seminal article on test reliability, Cronbach (1947) defines reliability of measurement as a property of the stability of performance over successive independent test administrations—a concept based on impossible assumptions of independence and constancy of successive behavior that “cannot be directly observed” (p. 2). Based on the classical true score model of reliability, measures of rater agreement attempt to quantify the variance within assessment situations that can be attributed to differences between raters. These coefficients are calculated as indices of the extent to which raters can reliably classify objects of measurement into specified scale categories. In rater-mediated writing assessments, agreement coefficients are used to assess the consistency of ratings across a set of essays. With a value of “1” and “-1” indicating perfect positive and negative correlations, respectively, and “0” indicating no linear relationship, higher absolute values of agreement coefficients are assumed to reflect higher similarity among ratings.

Numerous coefficients have been proposed to evaluate rater agreement based on assumptions and conditions that underlie specific measurement situations. Similarly, a variety of statistical indices of rater agreement exist across statistical software programs and methodological guides. Research on rater agreement demonstrates the application of many methods for examining the consistency of ratings in performance assessment. For example, Cohen’s kappa (Cohen, 1960), Kendall’s *w* (Kendall, 1939), coefficient alpha (Cronbach, 1947), and indices of dependability from Generalizability Theory (Brennan, 2001) can be found throughout literature on rater agreement. In addition, numerous intraclass correlation coefficients have been proposed as measures of reliability, with different versions applied depending on the assumptions and requirements of the measurement situation (Shrout & Fleiss, 1979). However, practical applications based on a traditional approach forgo most of these methods and routinely apply one or both of two simple measures of rater consistency to scores assigned by groups of raters: percent agreement and Pearson product-moment correlation coefficients (*r*) (Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991). Agreement indices based on a traditional approach can be calculated as a measure of the degree to which raters assign similar scores, i.e., the degree to which ratings are consistent within a specific measurement context. Percent exact rating agreement can be identified by cross-tabulating scores from pairs of raters in a table and examining the proportion of shared ratings (Johnson, Penny, & Gordon, 2009). In practice, cross-tabulations are used to identify consistency among pairs of operational raters, as well as among selected pairs of operational and expert raters (Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991). These measures are then used to identify raters who appear to be rating in a manner inconsistent with other raters and/or expert raters—this is viewed as a potential threat to rating quality.

Selection of an Agreement Coefficient for Rater-Mediated Writing Assessments

Different methods for assessing rater agreement yield different interpretations, and thus, different consequences for the use of scores. As a result of a wide variety of methods to estimate rater agreement, research on rater-mediated assessment has explored the relative usefulness of various estimation techniques across populations of students. For example, in their comparison of traditional indices of rater agreement with latent-class models, Schuster and Smith (2002) point out that there exists no single appropriate model as an index of rater agreement, and context-specific requirements must be used along with measures of model-data fit to select an appropriate method for the estimation of rater agreement. They emphasize that statistical models should be interpreted in terms of their specific properties, because individual models "typically represent specific perspectives or lenses on the data, exposing through their parameters focal properties of the phenomenon studied" (p. 392). Similarly, Zegers (1991) provides a useful overview of coefficients and their assumptions and implications for estimating interrater agreement. He reviews coefficients for nominal-level and non-nominal-level data and concludes that the selection of an agreement coefficients be informed by the implicit or explicit assumptions related to the nature of the data and for the specific assessment context.

Along with issues related to the broad range of methods for calculating interrater agreement, conceptual issues challenge the validity of agreement statistics as indices of quality. In their discussion of error and accuracy measures for performance appraisal ratings, Murphy and Cleveland (1991) describe an inherent conflict in the interpretation of interrater reliability measures. They discuss difficulty in determining the implications of rater agreement, stating: "It is not at all clear whether this criterion provides information about the reliability of ratings, the validity of ratings, or both," and note that disagreement among raters "cannot be attributed solely to random measurement error; different raters observe different aspects of the same ratee's performance and will sometimes honestly disagree in their evaluations" (p. 215). Similarly, Bond and Fox (2007) describe conceptual difficulty with the interpretation of indices of consistency as indices of quality. Instead, they call for consideration of student or rater characteristics that may result in consistent or inconsistent scores, stating: "If we reasonably harbor a suspicion that some attributes of the candidates are consistently important aspects of their performances, we could examine one or more facets related to candidates, such as gender, first language, cultural grouping, and the like" (p.147). These authors suggest that a simple measure of consistency among ratings may actually reflect agreement or disagreement on factors unrelated to the intended construct. As a result, indices of rating quality reported as agreement statistics may be insufficient measures of rating quality unless they are supplemented with additional information.

Rater Error and Systematic Bias

Although agreement coefficients are used to provide indices of rating quality across raters, their utility for demonstrating a clear view of rating quality is challenged by a lack of focus on specific trends and patterns that may indicate error or systematic bias. Based on the idea that chance or random error alone is not responsible for the variation in scores assigned by raters, studies that examine rater error and systematic bias seek to identify systematic variation that can be attributed to trends in rater behavior as they respond to performances. For example, research on achievement differences between groups of students on writing assessments suggests that characteristics such as gender (Cole, 1997), race/ethnicity (Engelhard, Gordon, Walker, & Gabrielson, 1994), or best language (Mattern, Camara, & Kobrin, 2007) influence rater perceptions of writing achievement, thus contributing to error and systematic bias.

A variety of definitions exist for rating patterns assumed to reflect biased or erroneous use of rubrics. As they are presented in the performance assessment literature, errors and systematic biases are aberrant patterns of rating scale use that contribute to the assignment of scores different from those hypothesized as true reflections of a student's achievement. An examination of the literature reveals a wide range of definitions for rating errors and biases, with an equally wide range of classification methodologies. Murphy and Cleveland (1991) present a useful review of literature related to the impact of rater errors and systematic biases on rating quality, and use four major categories to identify and describe rating errors: (1) severity/leniency, (2) halo, (3) response sets (e.g., central tendency), and (4) range restriction. Engelhard (2002) extends these categories to include interaction effects among facets in the measurement model and differential facet functioning (DFF) as a function of construct-irrelevant components. Detailed descriptions of these major categories are provided in Saal, Downey, and Lahey (1980) and Engelhard (1994; 2002).

Research on performance assessment, particularly in the area of writing assessment, has examined methods for identifying rater errors. As noted by Saal, Downey, and Lahey (1980), rater errors can be used as criteria to evaluate the psychometric quality of ratings. Various methods have been proposed for the identification and calibration of specific categories of error and systematic bias in ratings, as well as the evaluation of consistency within and across student populations and performance tasks.

Rater Error and Systematic Bias within a Traditional Approach

In practice, traditional measures of rating quality often use a comparison of rater means as an indicator of quality (Johnson, Penny, & Gordon, 2009). Rater severity and leniency can be easily identified by calculating and comparing mean scores assigned by individual and groups of raters. Engelhard (2002) defines severity and leniency error as a systematic "tendency on the part of raters to consistently provide ratings that are higher or lower than warranted by examinee performances" (p. 272). As demonstrated by Coffman (1971), raters whose scores tend to be significantly above the group mean for a particular student or domain-score area may be demonstrating leniency error, while raters whose scores are significantly below the group mean may be demonstrating severity error. Mean ratings provide a useful index of rater consistency that can be used to identify outliers and examine these scores for evidence of rater error or systematic bias. In addition to an examination of rater means across the entire group of students, mean scores can be compared for subgroups of students (e.g., gender or race/ethnicity groups) to identify errors or systematic bias related to student characteristics that may be contributing to differences in rating quality for these groups.

Rater Error and Systematic Bias within a Rasch-based Approach

The many-facet Rasch model (MFRM) can be applied as a method for identifying rater errors and systematic biases as indices of rating quality. Engelhard (1994) describes the application of the MFRM as a useful method for detecting and distinguishing among various types of rater error and systematic bias. Similarly, Kondo-Brown (2002) and Wolfe (2004) use Facets analyses (Linacre, 2010) in writing assessments to examine the range of rater severity/leniency as well as rating scale category use as indicators of rating quality. Congruent with Engelhard (1994), findings from both studies indicate the practicality of Rasch models to identify significant differences in the distribution of raters and scale categories that may indicate error or bias for groups of students.

The Rasch approach to identifying rater error and systematic bias compares the distribution of ratings and rating scale categories to expected distributions, and uses Infit and Outfit Mean Square Error (*MSE*) statistics to identify individual raters whose ratings differ significantly from those expected by the model. In his discussion of rater errors, which he calls “rater misbehavior,” Linacre (2010) describes methods for identifying rater severity and leniency within the Facets program. The Facets output provides a direct measure of rater severity that calibrates and maps individual raters in a visual display for easy comparison. This method transforms observed ratings to log-odds ratios on a common interval scale, with 0.00 often used to represent the mean severity measure (Bond & Fox, 2007). Raters can then be described in terms of their relative locations on the logit scale.

Fit statistics provide information about the amount of variability in data compared to the variability that would be expected based on an observation of perfect fit to the model (invariant measurement). These statistics provide information about the degree to which facets of the assessment situation diverge from the expectations of the Rasch model, and are useful for measuring the intended construct (Bond & Fox, 2007; Engelhard, 2002). Findings of “misfit” suggest that the observed data are not summarized well by the model. High values indicate data patterns that are more haphazard or “noisy” than expected, and low values indicate more uniform or “muted” patterns than were expected (i.e., rating patterns that appear “too good to be true”). In the case of raters, Infit and Outfit *MSE* statistics are used to identify rating patterns that are neither noisy nor muted, but are productive for measurement of writing achievement.

Infit and Outfit *MSE* statistics are useful within the context of rater-mediated assessments (Engelhard, 1994; Linacre, 2010). Outfit *MSE* statistics (u_j) and Infit *MSE* statistics (v_j) are provided in the output from Facets, and provide an index of unweighted mean square residual differences between observed and expected patterns in rating data. The Outfit *MSE* statistic for the rater facet is useful because it is particularly sensitive to "outliers," or extreme unexpected patterns in rating. Infit *MSE* statistics are also useful for evaluating model-data fit, but are less sensitive to outlying data. Outfit *MSE* is calculated by summing over the appropriate values and, as given in Engelhard (1994) is:

$$u_j = \sum_{n=1}^{N} \sum_{i=1}^{I} z_{nij}^2 / (N \times I), \tag{2}$$

and the formula for Infit is:

$$v_j = \sum_{n=1}^{N} \sum_{i=1}^{I} W_{nij} z_{nij}^2 \; / \; \sum_{n=1}^{N} \sum_{i=1}^{I} W_{nij} \, , \tag{3}$$

where

n = student,

i = domain,

j = rater,

k = number of rating scale categories,

and

$$W_{nij} = \sum_{x=0}^{m} (k - E_{nij})^2 \, \pi_{nijk} \, . \tag{4}$$

Standardized versions of these two fit statistics are calculated based on a z-score transformation, and they are useful for creating visual displays such as scatter plots because of their tendency towards linearity.

Engelhard (2002) demonstrates applications of Infit and Outfit *MSE* statistics to ratings from a writing assessment, and describes these two statistics as a method for determining the consistency of ratings within domains across groups of students. The mean square and standardized residual fit *MSE* statistics can be used to identify raters whose scores tend to differ significantly from those expected based on the MFRM. As in the traditional approach, fit statistics can be examined at the student subgroup-level for evidence of error or systematic bias that may indicate a threat to rating quality.

Interpreting Rater Error and Systematic Bias in Context

Research on rater error and systematic bias highlights a need for consistent definitions and identification methods for categories of error and bias. Noting different implications for rater errors when they are inconsistently defined, Saal, Downey, and Lahey (1980) describe a “lack of congruency between conceptualization and quantification” of these phenomena (p. 423). In their discussion of psychometric characteristics of ratings, Murphy and Cleveland (1991) describe rater errors as indirect indicators of rating quality, and note inconsistency between the definitions of rating errors and methods for identifying them (e.g., different units of analysis). Lack of correlation among rating errors and direct measures of rating accuracy, they claim, is evidence that traditionally defined rating errors do not necessarily suggest decreased quality of ratings.

Rater Accuracy

Rater accuracy constitutes a third category of rating quality indices, and the concept of accuracy has been used both directly and indirectly to evaluate the quality of ratings in performance assessment. Wolfe and McVay (2011) describe rating inaccuracy as an effect of “seemingly random error that causes the assigned ratings to be inconsistent with accurate ratings” (p. 6). Much research on ratings in performance assessment uses indices of rater agreement and errors as indirect indices of rater accuracy, associating high levels of agreement and a lack of errors with high quality. Direct estimates of accuracy are based on the match between operational ratings and those established as true ratings by individual or committees of "expert" raters. In their review of research on measures of rating accuracy, Sulsky and Balzer (1988) describe the process for estimating rater accuracy as a comparison of "the rater's performance evaluations for *n* ratees on *k* performance dimensions with corresponding evaluations provided by 'expert raters'" (p. 498). This conceptualization of accuracy as a function of the difference between operational and expert ratings is evident throughout research on rating quality, as demonstrated by the accuracy estimation procedures discussed below.

In practice, methods for estimating rater accuracy to evaluate the quality of ratings are used within both traditional and Rasch-based approaches. Traditional approaches for the assessment of rater accuracy include the use of percent accuracy agreement, or the D^2 index for accuracy (Cronbach, 1955), and measures of correlational accuracy; modern approaches estimate severity or leniency accuracy, and explore rater accuracy using fit statistics (Engelhard, 1996).

Accuracy Measures within a Traditional Approach

Consistent with Sulsky and Balzer's (1988) definition, the most common procedure for estimating accuracy in performance assessment literature involves an estimation of the distance between operational and "true" or "expert" ratings. An early method developed for the estimation of rating accuracy is percent accuracy agreement. This measure is equivalent to the D^2 statistic and measures the distance between operational and true ratings averaged across ratees and dimensions, or domains in an analytic rubric (Sulsky & Balzer, 1988, p. 498). D^2 can be expressed mathematically as:

$$D^2 = \frac{1}{kn} \sum \sum_{nk} (x_{nk} - t_{nk})^2, \qquad (5)$$

where

x = subject ratings,

t = true scores,

n = students, and

k = dimensions.

This measure of accuracy uses ratings from “expert” raters as criteria against which to compare scores assigned by operational raters. Johnson, Penny, and Gordon (2009) describe percent accuracy agreement for rater monitoring as a “validity check,” based on the premise that congruence between operational and expert ratings provides evidence of rating quality. Contrasting accuracy agreement with other measures of rating quality, they claim that this index “introduces the validity of rater scores that is absent from measures of interrater agreement” (p. 235). They describe the usefulness of computer-based rating systems, in which percent accuracy agreement can be continuously monitored, for monitoring and evaluating rating quality from large numbers individual and groups of raters throughout the scoring process.

Within a traditional approach, statistical techniques have been developed that separate the percent accuracy agreement index into individual components of accuracy. For example, Cronbach (1955) decomposes the distance between operational and true ratings into four unique sources based on ANOVA that communicate distinct information about aspects of the difference between and correlations among operational and expert ratings: elevation accuracy (E), differential elevation (DE), stereotype accuracy (SA), and differential accuracy (DA). However, observing that each of the four accuracy components provides unique information, Cronbach (1955) emphasizes an absence of necessary correlation between the four scores, and cautions against the use of a single accuracy component score to make general assumptions about overall rater accuracy. In practice, traditionally based accuracy measures as indices of rating quality most often rely on percent accuracy agreement statistics as a single estimate of the match between operational ratings and those assigned by "experts" (Berkowitz-Jones, 2007; Johnson, Penny, & Gordon, 2009).

Accuracy Measures within a Rasch-based Approach

The MFRM can be used to evaluate the accuracy of operational raters on a dichotomous accuracy scale, based on a match between operational and expert ratings on benchmark papers, with "0" reflecting a lack of equivalent scores, or an inaccurate rating, and "1" reflecting a perfect match, or an accurate rating (Engelhard, 1996). The Facets model for dichotomous data can be applied, which is expressed mathematically as:

$$ln\left[\frac{P_{ni1}}{P_{ni0}}\right] = \beta_n - \delta_i \,, \tag{6}$$

where

P_{ni1}= probability of rater *n* being accurate (x = 1) on paper *i*,

P_{ni0} = probability of rater *n* being inaccurate (x = 0) on paper *i*,

β_n= accuracy of rater *n,* and

δ_i= difficulty of being accurate on paper *i*.

Similar to the Rasch-based approach to evaluating rater severity in terms of error and systematic bias, the MFRM can be used to calibrate individual raters in terms of their severity/leniency accuracy, allowing for a direct comparison of a set of raters on this measure. When the dichotomous accuracy model is applied, the Facets (Linacre, 2010) output calibrates and maps raters in terms of their accuracy. An examination of the relative placement of raters on the logit scale can then be used to evaluate the severity and leniency accuracy of individual raters.

Along the same lines, fit statistics can be used to evaluate accuracy score patterns for individual raters. The mean square and standardized residual fit statistics can be used to flag raters whose accuracy scores tend to differ significantly from those expected by the model. Similar to their use for identifying rater error and systematic bias, fit statistics for accuracy scores can be examined at the student subgroup level for evidence of patterns that may indicate a threat to rating quality.

The use of fit statistics as a measure of rater accuracy is demonstrated by Engelhard (2007), who applied the MFRM to writing assessment data in order to detect what he terms *differential rater functioning* (DRF). This approach to evaluating rating quality uses residual analyses to flag raters whose scores differ significantly from expected ratings. In his study, DRF was compared for male and female students to identify differences in rater accuracy as a function of construct-irrelevant influences on scores by gender. DRF was found in this subgroup analysis, highlighting the need for "detailed examination of residuals for each rater" followed by mixed-methods investigations of "suspect raters" with regard to differential functioning based on construct-irrelevant factors such as gender (Engelhard, 2007, p. 1124).

Interpreting Rater Accuracy in Context

Selection of a method for estimating rater accuracy to evaluate the quality of ratings depends largely on the assessment context. As described in Murphy and Cleveland (1991) and Sulsky and Balzer (1988), many approaches exist within psychometric methods for estimating rater accuracy, including accuracy measures based on Cronbach's (1955) components of accuracy and Borman's (1977) measure of differential accuracy. Sulsky and Balzer (1988) point out that measures of rater accuracy "require the direct comparison of ratings obtained from a single rater with true scores to compute an accuracy index" (p. 500). The availability of a single "expert" rating from which to compute accuracy scores is not always practical in operational settings. Furthermore, as noted by Cronbach (1955, 1958), Sulsky and Balzer (1988), Murphy and Cleveland (1991) and others, the various methods for computing rater accuracy provide inconsistent and uncorrelated results—highlighting the necessity for a contextualized interpretation of accuracy measures.

Methodology

The above discussion addressed the first guiding question in this study, which sought to identify major categories for indices of rating quality in rater-mediated performance assessments of writing. In this section, an illustrative data analysis is presented as a method to address the second guiding question in this study: Do indices of rating quality within a traditional and Rasch-based approach identify raters in the same way in terms of rating quality?

This is a secondary analysis of data that was previously examined by Gyagenda and Engelhard (2009) who explored differences among domain-level scores for male and female students in the Georgia High School Writing Test, and by Andrich (2010), who presented findings of a structural halo effect across domains for raters. The analyses presented in this section apply evaluative techniques from indices of rating quality within a traditional approach as well as techniques from indices based on Rasch Measurement Theory in order to compare the two approaches and explore their respective implications within the context of a statewide rater-mediated writing assessment.

Instrument

Data used in this study were collected during a pilot administration of the Georgia High School Writing Test. This assessment requires students to respond to a persuasive essay prompt in a one- or two-page essay. During the administration used to collect these data, students were given 90 minutes to complete the assessment, and compositions were rated analytically on a four-point scale within four separate domains. Regarding the meaning of their writing, students were scored in terms of style and organization; regarding the mechanics of their writing, students were scored in terms of conventions and sentence formation. In this study, domains and rating categories are considered in terms of the definitions set forth in the (1993) Georgia High School Writing Test Assessment and Instructional Guide (Georgia Department of Education, 1993). Each point value represents a rating between one and four, with a score of one as "inadequate," two as "poor," three as "good," and four as "very good." Operational definitions for domains are presented in Table 1.

Participants

The data in this study include observations in the form of scores from essays composed by 365 eighth-grade students in Georgia, rated by 20 operational raters, and assigned a single score by a "validity" committee of six to eight "expert" raters. The 20 operational raters were randomly selected from a group of 87 raters hired to score essays for this administration of the assessment. Each selected operational rater and the validity committee scored all 365 essays. Prior to scoring, raters were required to complete a training program at the Georgia Center for Assessment, which involved instruction related to the rubric, prompt, and scoring practices specific to the Georgia High School Writing Test. In order to successfully complete training, raters were required to complete the 14-hour program and earn a passing scores on a qualifying test designed to assess their use of the rubric in terms of the specific requirements and intentions for prompts used in the 1993 administration. In order to maintain confidentiality, all identifying information about students, raters, schools, and districts was stripped from the data, and replaced by non-traceable identification numbers for students and raters. Information about schools and districts was completely removed from the data. IRB documentation for this study is provided in Appendix A.

Data Analysis

The data analysis for this study involved several steps. Based on the literature review, seven indices of rating quality were selected to illustrate relationships among indices of rating quality within a traditional and Rasch-based approach. These seven indices, summarized in Table 2: percent exact rating agreement, rater means, and percent exact accuracy agreement within the traditional approach, and rater severity measure, fit statistics for ratings, rater accuracy measure, and fit statistics for accuracy within the Rasch-based approach. Next, these statistics and measures were computed and compared. Third, Pearson product-moment correlations were computed across indices within the two approaches in order to examine their relationships. Graphical displays were also examined for these data.

Before beginning analyses, it was necessary to transform ratings to accuracy scores in order to examine rater accuracy as an indicator of rating quality. A comparison of ratings with those assigned by the validity committee was used to create a second data set of dichotomous accuracy scores. Ratings that were in exact agreement with those assigned by the validity committee were considered accurate, and those that were different were considered inaccurate. For the rater accuracy analyses in this study, ratings that matched the validity committee were assigned an accuracy score of '1,' and ratings that were discrepant from the validity committee were assigned an accuracy score of '0,' such that higher accuracy ratings reflect more accurate scoring in terms of a match between operational rater and validity committee scores. The dichotomous accuracy ratings were then used to conduct accuracy analyses.

Selected Indices within a Traditional Approach

As shown in Table 2, three indices were selected based on the traditional approach to evaluating ratings: percent exact rating agreement, rater means as a measure of error and systematic bias, and percent accuracy score agreement as a measure of rater accuracy.

The traditional approach to evaluating ratings in performance assessment associates high levels of agreement among ratings with high quality (Murphy & Cleveland, 1991). In this study, rater agreement within the traditional approach was examined using a percent exact agreement statistic. When the rating-scale model is used, the “inter-rater=” command in Facets (Linacre, 2010) gives a measure of inter-rater agreement that is the percentage of identical ratings assigned in the same circumstance (in this case, individual student compositions) by different raters (Linacre, 2010). By examining the percent of identical ratings assigned by a group of raters, it is possible to investigate the degree to which raters are making unique and consistent judgments about a group of students (Linacre, 2010).

Although several types of rater error and systematic biases are discussed in the performance assessment literature (see, for example, Engelhard, 1994), this study focuses on the rater error of severity in order to illustrate the relationships among rating quality indices within a traditional approach. In order to identify raters whose scoring patterns tend to be significantly above or below the mean, rater means across all 365 essays were computed for each of the operational raters, and the raters were compared in order to identify significantly severe or lenient raters in this sample. This method is commonly used in practice to identify severe and lenient raters (Johnson, Penny, & Gordon, 2009; Coffman, 1971).

Rater accuracy was examined within the traditional approach using a similar method to the traditional measure of rater agreement. The traditional approach to evaluating rater accuracy used in this study involved computing an index of percent exact accuracy score agreement for each of the 20 operational raters. This statistic is an indicator of how often rater accuracy scores were identical in the same circumstance (in this case, on the same essay). As with percent exact agreement among ratings, this index of accuracy score agreement assumes that higher similarity among rater accuracy scores is associated with higher quality ratings.

Selected Indices within a Rasch-based Approach

This study uses four indices of rating quality based on Rasch Measurement Theory. Based on invariant measurement, the MFRM used here accounts for variability in scores as a combination of student writing achievement, rater severity, and domain difficulty. As a result, differences among raters are captured in estimates of rater severity, and fluctuation in rater judgment is captured in Infit and Outfit *MSE* statistics. Therefore, within the Rasch measurement theory framework, raters are viewed as "individual experts," and rater agreement was not considered relevant within this approach. Instead, Rasch-based analyses were limited to indices of rating quality within the categories of rater error and systematic bias and rater accuracy.

Rater error and accuracy analyses were conducted using Facets (Linacre, 2010) to calibrate each of the operational raters in terms of severity and accuracy, and to compare Infit and Outfit *MSE* statistics in order to examine model-data fit. The Rasch-based calibration of rater severity and accuracy provides a straightforward method for comparing rating and accuracy score patterns in this context. Raters whose rating patterns tended to be more severe were located higher on the logit scale, and raters with more lenient rating patterns were located lower on the logit scale. Similarly, raters whose accuracy scores tended to be higher were placed higher on the logit scale (indicating higher accuracy "ability"), and those with lower accuracy scores were placed lower on the logit scale.

Fit statistics were also used within the Rasch-based approach in order to examine model-data fit for ratings and accuracy scores. Fit statistics allowed raters to be identified whose rating patterns varied more or less than was expected by the model. Raters with high values of fit statistics were assumed to be exhibiting rating patterns that were more haphazard or "noisy," than expected, and raters with low values of fit statistics were assumed to be exhibiting overly-predictable, or "muted" rating patterns. Raters with fit statistics around 1.0 are considered to be exhibiting rating patterns close to those expected by the model, thus providing measures that can be assumed to yield useful information about student achievement in terms of the construct.

Results

Summaries of findings from the data analyses are presented in Tables 3-11. Results from this study are described in four sections. First, summaries of statistical and psychometric measures for ratings and accuracy scores, along with domain calibrations and rating scale category information are presented, followed by results from separate analyses of rater agreement, error and systematic bias, and accuracy for this sample of raters. Third, results are described from the comparison of the three traditional rating quality indices. Finally, results from the comparison of rating quality indices based on Rasch measurement theory are presented.

Summary of Statistical and Psychometric Measures

Calibration of Ratings

Table 3 provides summary statistics from Facets for the student, rater, and domain calibrations in this study based on the rating data. As shown in the table, the overall differences between students (θ), raters (λ), and domains (δ) are significant ($p < 0.05$), with high reliabilities of separation ($REL_\theta = 0.99$; $REL_\lambda = 0.98$; $REL_\delta > 0.99$). Good fit to the model is evident for each of these three facets, with mean Infit and Outfit statistics around 1.00, and standard deviations around 0.20. Acceptable model-data fit suggests that the MFRM is functioning as intended for these rating data. The variable map shown in Figure 2 presents a graphical display of the spread of student, rater, and domain measures for this sample. Detailed calibration information for the 20 operational raters is given in Table 4.

Domain Calibrations for Ratings

The calibration of the domain facet for ratings is presented in Table 5, with summary statistics provided in Table 3. The difficulties of the domains range from -0.92 logits (*SE* = 0.02) for sentence formation to 0.68 logits (*SE* = 0.02) for organization. As shown in Table 3, the overall difference between the four domains is significant (χ^2 (3, N = 4) = 2654.2, $p < 0.05$) with a high reliability of separation index ($REL_\delta > 0.99$). There is no evidence that the raters are using the domains in an inconsistent fashion, and all Outfit and Infit *MSE* statistics are between 0.93 and 1.05, indicating appropriate fit to the Facets model for these rating data. The ordering of domain difficulty in these data is congruent with what may be expected in terms of the construct. Aspects of the writing process that require less abstraction, or lower-level thinking, are likely reflected in domains related to mechanics—estimated here as less difficult; abstraction, or higher-order thinking is likely reflected in domains related to meaning—estimated here as more difficult.

Rating Scale Category Use

Table 6 summarizes operational rater use of the rating scale for this sample. Inspection of the rating scale structure for the observed scores indicates that there is a generally good spread of ratings across scale categories. In addition, it appears that the operational raters in this sample are using rating scale categories as intended, evidenced by the order of category placement on the logit scale. Category 1, which represents an “inadequate” response, has the lowest measure (-3.97 logits), suggesting that it is easiest, and Category 4, which represents a “very good” response has the highest measure (4.15 logits), suggesting that it is most difficult.

Calibration of Accuracy Scores

Table 7 provides summary statistics from Facets for the rater (β), benchmark paper (δ), and domain calibrations (α) related to the accuracy analyses in this study. The rater facet represents the spread of rater accuracy measures based on the dichotomous scale. The benchmark paper and domain facets represent the spread of difficulty for raters to be accurate on a particular paper or within a particular rubric domain. As shown in the table, the overall differences between rater accuracy calibrations (β), benchmark paper difficulty for accuracy (δ), and domain difficulty for accuracy (α) are significant ($p < 0.05$), with high reliabilities of separation (REL_{β} = 0.97; REL_{δ} = 0.83; REL_{α} = 0.84). Good fit to the model is evident for each of these three facets, with mean Infit and Outfit *MSE* statistics around 1.00, and standard deviations around 0.20. Acceptable model-data fit suggests that the dichotomous Rasch model is functioning as intended for these accuracy score data. The variable map in Figure 3 provides a visual display of the spread of these three accuracy facets for this sample. Detailed calibration information for the operational rater accuracy scores is given in Table 8.

Domain Calibrations for Accuracy Ratings

The descriptive statistics for the accuracy calibration of the four domains are presented in Table 9. The difficulties of the domains for rater accuracy range from -0.07 logits for sentence formation (*SE* = 0.03), which is the easiest domain for operational raters to rate accurately, to 0.06 logits for Meaning/Style (*SE* = 0.03), which is the hardest domain for operational raters to rate accurately. Although the visual display in Figure 3 appears to indicate similar measures for domain difficulty, Chi-Square statistics indicate that the differences between rater accuracy on the four domains are statistically significant (χ^2 (19, N = 20) = 326.90, $p < 0.05$) with a reliability of separation index of $REL_\alpha = 0.84$. As shown in the table, the two mechanics domains (sentence formation and conventions) were easier for raters to score accurately than were the two meaning domains (organization and style). This ordering of domains suggests that mechanics domains, which are often considered less subjective, are more often scored accurately than are domains related to the meaning of writing; the order of difficulty matches the order that may be expected in terms of the construct. In addition, Outfit and Infit *MSE* statistics for the domains ranged from 0.98 to 1.02, suggesting good fit to the model for the accuracy data.

Rating Quality Analyses

Rater Agreement

Rater agreement as an indicator of rating quality was examined within the traditional approach by computing percentages of exact rating agreement in Facets (Linacre, 2010). Table 4 presents this measure of agreement for each of the 20 operational raters in the sample. For this sample, rater agreement ranges from 54.3% (Rater 18) to 61.7% (Rater 4), indicating a relatively small range of variation. These agreement statistics indicate that, on average, the raters in the sample agreed with one another about 58% of the time. Rater agreement was not examined using measures from Rasch measurement theory.

Rater Error and Systematic Bias

The traditional approach for identifying rater error and systematic bias involved a comparison of mean ratings across the sample of operational raters to identify those who appeared to be rating severely or leniently. In order to identify operational raters whose ratings tended to be substantially higher or lower than the rest of the group, a mean rating was computed for each of the 20 operational raters (given in Table 4). Within a traditional perspective, large variation among raters is thought to indicate potential threats to rating quality, with raters whose mean scores are significantly above the rest of the group demonstrating an error of severity, and those with significantly lower scores demonstrating leniency. As can be seen in Table 4, there is not much variation among mean ratings, with the most severe rater (Rater 9, *M* = 2.86) is about 0.15 points above the group mean (*M* = 2.71, *SD* = 0.09), and the most lenient rater (Rater 15, *M* = 2.52) about 0.19 points below the group mean.

Analyses based on Rasch measurement theory to identify rater error and systematic bias were conducted in the Facets program and included rater calibration measures on the logit scale as an indicator of severity, and fit *MSE* statistics to identify raters whose rating patterns were either highly fluctuating (“noisy”), or overly predictable (“muted”). The Rasch-based severity calibration provided a straightforward method for comparing the severity of the 20 operational raters in this sample. As shown in Table 4, rater measures of severity ranged from -0.59 logits (*SE* = 0.05) for Rater 9 who is lenient to 0.76 logits (*SE* = 0.05) for Rater 15 who is severe. The overall difference between rater severity measures are significant (χ^2 (19), *N* = 20) = 869.5, *p* < *.05*) with a high reliability of separation index (REL_{λ} = 0.98). The spread of Infit and Outfit *MSE* statistics for ratings indicated acceptable fit to the model, with the lowest Outfit value observed for Rater 17 (Outfit *MSE* = 0.78) and the highest value observed for Rater 18 (Outfit *MSE* = 1.25). The lowest Infit *MSE* value was observed for Rater 17 (Infit *MSE* = 0.78) and the highest value for Rater 6 (Infit *MSE* = 0.78).

Rater Accuracy

Percent exact accuracy score agreement values among the 20 operational raters are given in Table 8, and range from 57.50 % agreement for Rater 15 who is least accurate, to 63.00% agreement for Rater 19, who is most accurate by this index.

Similar to the Rasch-based approach for evaluating rater error and systematic bias, the Rasch-based approach for assessing rater accuracy included the calibration of individual raters on the logit scale and the use of Outfit and Infit *MSE* statistics to identify unexpected patterns in rater accuracy scores. Results from these analyses are presented in Table 8. Measures of rater accuracy on the logit scale range from -0.37 logits (*SE* = 0.06) for Rater 19, who is the least accurate operational rater in the sample, to 0.85 logits for Rater 10 (*SE* = 0.05), who is the most accurate. The overall difference between rater accuracy measures is significant (χ^2 (19, *N* = 20) = 622.6, $p < 0.05$) with a high reliability of separation index ($REL_\beta = 0.97$). The spread of Infit and Outfit *MSE* statistics for accuracy ratings indicated acceptable fit to the model, with the lowest value observed for Rater 11 (Outfit *MSE* = 0.93; Infit *MSE* = 0.96) and the highest value observed for Rater 3 (Infit *MSE* = 1.03; Outfit *MSE* = 1.71).

Comparison of Rating Quality Indices

In order to answer the second research question in this study, Pearson product-moment correlations (*r*) were computed in SPSS across the indices of rating quality within the traditional and Rasch-based approach. The value of this coefficient can range from negative one to positive one and represents the strength and direction of a linear relationship between two variables. Values that approach positive and negative one reflect strong linear relationships. Values of the correlation between indicators for the 20 raters in this sample are given in Tables 10 and 11, and scatter plots for significant across-category correlations are shown in Figures 4 and 5. Because a large number of comparisons were used for this correlational analysis, the Bonferroni correction procedures was applied, and adjusted *p* values were used to judge the significance of correlations. The adjusted values used for significance testing are given in the Tables. First, correlations within the traditional approach are discussed, followed by a discussion of findings within the Rasch-based approach.

Table 10 shows Pearson product-moment correlations between the traditional indices of rating quality: percent exact rating agreement, rater means, and percent exact accuracy score agreement. Scatter plots for these three correlations are given in Figure 4. If Bonferroni-corrected alpha levels are used to control for the inflated risk of Type I error that occurs with the use of multiple significance tests, the alpha level is lowered to 0.01/3 = 0.0033 for this set of correlations. Non-significant correlations were found between percent exact rating agreement and rater mean (*r* (18) = -0.0850, *p* = -0.7216, Panel A), and between percent exact accuracy score agreement and rater mean (*r* (18) = -0.1277, *p* = .5916, Panel B). These non-significant correlations suggest a lack of alignment between the indicator of rater error (rater mean) and the indices of rater agreement and rater accuracy within the traditional approach. In contrast, a significant correlation was found between percent exact rating agreement and percent accuracy agreement (*r* (18) = 0.939, *p* = 0.0000, Panel C). Although these two statistics are both related to agreement, the strong correlation is not a statistical artifact. The two statistics indicate different properties of the data: percent exact rating agreement represents the match between an operational rater's rating on a particular paper with the rest of the ratings assigned to that paper, and percent exact accuracy score agreement represents the match between rater accuracy scores on individual papers, which are based on ratings from the VC. This significant correlation suggests congruence between indicators of rater agreement and rater accuracy based on a traditional approach for these data.

Table 11 and Figure 5 show correlations between the indices of rating quality based on Rasch measurement theory that were selected for this study: rater severity calibration, model-data fit for ratings, rater accuracy calibration, and model-data fit for accuracy. In order to present a clear visual display, the standardized version of the Outfit *MSE* statistic is the only indicator of model-data fit depicted in Figure 5 because of its likelihood to be linear. If Bonferroni-corrected alpha levels are used to control for the inflated risk of Type I error that occurs with the use of multiple significance tests, the alpha level is lowered to 0.01/45= 0.0002 for this set of correlations. Four of the significant correlations in this table are a result of the transformation between the two fit statistics and their respective standardized versions. Exempting these four transformations as artifacts due to the way they are calculated, non-significant correlations were found between all of the indicators of rater error and accuracy except for a significant correlation observed between the direct measure of rater accuracy and Outfit *MSE* for ratings (r (18) = -0.743, p = 0.0002), and between the direct measure of rater accuracy and the standardized Outfit statistic for ratings (r (18) = -0.766, p = 0.0001). These significant correlations suggest alignment between rater error in terms of model-data fit, and the direct measure of rater accuracy based on Rasch measurement theory. This finding suggests that indices of model-data fit may be a possible indicator of rater accuracy in the context of rater-mediated writing assessment.

Discussion

The purpose of this study was to explore the consistency among different indices of rating quality within the context of a large-scale rater-mediated writing assessment. Data from studies by Gyagenda and Engelhard (2009) and Andrich (2010) were used in a secondary analysis to determine the degree to which measures of rater agreement, error and systematic bias, and accuracy are consistent within a traditional and Rasch-based approach. Findings of significant and non-significant correlations among these three categories of indices hold significant implications for evaluating the quality of ratings in rater-mediated assessments—an area of growing importance in assessment research.

Several conclusions can be drawn from these findings. The literature review revealed a variety of indices that can be used to evaluate the quality of ratings in large-scale rater-mediated writing assessments. These indices can be grouped into three major categories: (a) rater agreement, (b) rater error and systematic bias and (c) rater accuracy. However, the definitions and applications of indices within each category varied widely. In general, performance assessment literature indicates a trend of inconsistency in approaches to the evaluation of rating quality, and, as a result, a general dissatisfaction with ratings as measures of student achievement in high-stakes contexts. Landy and Farr (1980) conducted a literature review on ratings in performance assessment and concluded that performance rating is "best thought of as a system comprising many different classes of variables" that are "vulnerable to both intentional and inadvertent bias" (p. 73). In response to widespread reservations for the use of ratings to evaluate student achievement, Saal, Downey, and Lahey (1980) point out that the evaluation of rating quality, and thus, the quality of scores, depends largely on the selection and definition of evaluative criteria for ratings. Calling for increased precision in the methods and definitions of indices used to evaluate ratings, they assert the fact that, "since the choice of operational definitions of the psychometric qualities of rating data can actually determine the results of such a comparative study, that choice must be something more than an arbitrary exercise governed by choice" (p. 421).

The empirical data analysis revealed several trends related to the alignment between rating quality indices within the traditional and Rasch-based approach. Correlations among traditional indices of rating quality indicated a lack of alignment between measures of rater error and rater agreement, and between rater error and rater accuracy. However, a strong and significant correlation between the traditional indicators of rater agreement and rater accuracy was found. This suggests that these two measures within the traditional approach provide similar information about a set of raters in terms of rating quality.

Within the Rasch-based approach, non-significant correlations were found among all of the indicators of rating quality except between Outfit *MSE* statistics used to evaluate model-data fit for rating patterns and the direct measure of rater accuracy on the logit scale. This finding is particularly significant for rater-mediated performance assessment. Previous research on performance assessment has not found an alignment between indicators of rater error and direct measures of rater accuracy (Landy & Farr, 1980; Murphy & Cleveland, 1991; Sulsky & Balzer, 1988). Based on this finding, it may be possible to use measures of model-data fit as an indicator of rater accuracy for writing assessments. Because practical limitations during the scoring of writing assessments do not often permit scoring by a validity committee or single "expert" rater, the link found here between Rasch-based estimates of model-data fit and direct accuracy measures holds significant implications as an efficient method for evaluating rater accuracy in the context of writing assessment.

Limitations and Delimitations

The inferences made in this study apply to the sample of raters, and generalize to other trained raters or rater groups who score large-scale writing assessments. The model used in this study is a fixed-effects model, and the invariance assumption from Rasch measurement theory applies within the group of raters used in the study.

The sample size of the rater group limits the statistical power for the non-significant correlations found in this study (Cohen, 1988). However, the Pearson product-moment correlations between pairs of rating quality indices were cross-checked with Spearman (nonparametric) correlations in SPSS, and the same correlations were flagged as significant and non-significant. In addition, adjusted significance values from the Bonferroni correction procedure did not suggest violation of Type I error. As a result, the more-traditionally-used Pearson product-moment correlation index was preferred to a nonparametric coefficient.

It is also important to note that the scores from the validity committee used in these data to calculate operational rater accuracy scores were accepted for these analyses as consistent. However, the application of the rating scale by the committee may have shifted during scoring these 365 essays as a result of fatigue or other factors related to group dynamics. Moreover, the information provided in the accuracy analyses presented here is limited by the fact that accuracy scores were computed on a dichotomous scale and do not reflect the magnitude of difference between operational and VC scores. For example, this study treats a score difference of a single point as equal in magnitude to a score difference of three points. The possibility that the magnitude of difference between operational and VC scores plays a role in the correlation between model-data fit and accuracy measures should be examined in further studies.

Delimitations for this study are related to the selection of the research questions and data set. This study is guided by two questions based on three major categories of rating quality indices found in psychometric and performance appraisal literature on rater-mediated performance assessment. Selected indices of rater agreement, error and systematic bias, and accuracy were used in this study because they are commonly discussed and applied in practice for the evaluation of the quality of rater-mediated performance assessment. However, methods other than those used in this study exist within each of these categories, and, if applied and compared in a similar fashion as was done here, may result in different findings. For example, some methods for evaluating rater agreement that are not examined in this study include intraclass correlation coefficients (ICCs), Cohen's kappa coefficient (κ), or Cronbach's alpha, and measures of reliability from Generalizability Theory (Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991; Shrout & Fleiss, 1979). Similarly, other measures of rater error and systematic bias exist beyond severity and leniency, including halo error, error of central tendency, restriction of range, interaction effects, and differential facet (e.g., raters) functioning (Engelhard, 2002). Accuracy measures not used in this study include Cronbach's components of accuracy.

The data used in this study are from a pilot administration of Georgia's High School Writing Test administered to eighth-grade students in 1993. Since the administration of this pilot test, changes have been made to the statewide writing test for eighth-grade students in Georgia. However, because the rater training and scoring procedures as well as the relative emphasis on writing in Georgia's curriculum has remained consistent between the time of data collection and this study, the applicability of these findings is likely to remain significant. Furthermore, the study can be viewed as a methodological illustration for evaluating the quality of ratings in a large-scale rater-mediated writing assessment.

Conclusions in terms of Research Questions

Research Question One

What are the major indices of rater agreement, error and systematic bias, and accuracy that can be used to evaluate the quality of ratings in rater-mediated assessment?

In order to answer this question, performance assessment and writing assessment literature based on a traditional approach and on Rasch measurement theory was reviewed. The literature review revealed a variety of rating quality indices from both approaches that could be grouped into three major categories: rater agreement, rater error and systematic bias, and rater accuracy.

Rater agreement indices were based on a traditional approach. Rater agreement was interpreted as the degree to which raters make identical decisions regarding the achievement of students according to assessment criteria, and higher levels of agreement are interpreted as indices of high quality ratings. When rater agreement is used in practice to evaluate rating quality, discrepancies among ratings (lack of agreement) are attributed to random variation that occurs during the assessment process. Although a variety of rater agreement statistics were identified in the literature, a direct measure of the percent of identical ratings was found to be most frequently applied in practice (Johnson, Penny, & Gordon, 2009).

Indices of rater error and systematic bias were found in traditional and Rasch-based literature. Unlike rater agreement, indices of rater error are conceptualized within both approaches as indices of variation that is systematic, i.e., variation that can be linked explicitly to directional trends in rater behavior related to construct-irrelevant characteristics of raters, students, scoring procedures, or other aspects of an assessment process. Because rater errors and systematic biases are assumed to contribute to the assignment of scores that differ from those warranted by performance, they are interpreted within both the traditional and Rasch-based approach as detrimental to the quality of ratings. A variety of specific rater errors were found in the traditional and Rasch-based literature that are assumed to indicate different types of construct-irrelevant biases, thus limiting the quality of scores.

Indices of rater accuracy were found in the traditional and Rasch-based literature on rater-mediated assessment. Within both approaches, rater accuracy was defined as a match between a single "expert" rating and those assigned by operational raters. The evaluation of rating quality through rater accuracy indices is based on the assumption that expert ratings reflect the closest-possible approximation of model-defined true scores for student performance according to the rubric used for an assessment. Similar to rater agreement, high measures of rater accuracy are associated with high-quality ratings. In practice, rater accuracy is often assessed by interspersing "benchmark papers" that are scored by experts in order to monitor rating quality (Johnson, Penny, & Gordon, 2009).

In general, research on the correspondence among indices of rating quality found that indices of rating quality within these three categories were discrepant. Dissatisfaction with the wide array of definitions and methods for evaluating rating quality using these indices was reported in meta-analytic studies on rating quality indices by Landy and Farr (1980), Murphy and Cleveland (1991), and Sulsky and Balzer (1988).

Research Question Two

Do indices of rating quality from a traditional and Rasch-based approach identify raters in the same way in terms of rating quality?

Based on the review of literature used to answer the first research question, seven indices of rating quality were selected based on a traditional approach and Rasch measurement theory. The selected indices were then used in an empirical data analysis to evaluate the congruence among methods for evaluating rating quality within the context of a rater-mediated writing assessment. Rater agreement was evaluated by computing the percent of identical ratings on each of the student papers for the 20 operational raters in the sample. Rater error and systematic bias was evaluated using a comparison of mean ratings within the traditional approach, and a calibration of rater severity, and evaluation of model-data fit for ratings within the Rasch-based approach. Rater accuracy analyses were based on dichotomous accuracy scores (0 = inaccurate, 1 = accurate) and included a comparison of percent exact agreement among operational raters and scores assigned by a validity committee of expert raters within the traditional approach and rater accuracy calibration and evaluation of model-data fit for accuracy scores within the Rasch-based approach.

Separate analyses of rating quality based on each of these seven indices suggested that, in general, this sample of raters assigned high-quality ratings to the 365 student papers used in this study as defined by both approaches. The correlations between the seven indices revealed that discrepant information is provided by nearly all of the methods for evaluating rating quality that were examined in this study. However, two significant correlations were found in these data between the direct measures of rater accuracy and rater agreement within the traditional approach, and rater error within the Rasch-based approach. These significant correlations contrast with previous findings in rater-mediated assessment literature, which suggest that indices of rating quality are not comparable (Murphy & Cleveland, 1991; Sulsky & Balzer, 1988). These findings hold substantial implications for theory, research, policy and practice, which are discussed in the following section.

Implications

Theory

Findings of congruent rating quality indices will inform measurement theory related to rater-mediated assessment. The theoretical framework for this study was based on the idea that raters are a type of "lens" through which student achievement is made visible in an assessment context. The judgmental processes of raters can then be evaluated in terms of the influence of construct-irrelevant factors that may impact the scores assigned to student performance. Findings from this study revealed that numerous methods can be used to evaluate the degree to which construct-irrelevant factors influence rater judgment of student achievement. Because the type of information that is provided about a set of raters varies when different indices of rating quality are used to evaluate a group of raters, information about the variables that influence rater perception of student achievement also fluctuates. In other words, an awareness of the specific aspects of the judgmental processes, or the "lens" through which a rater judges student performance, must inform decisions about the indices of rating quality that are selected to evaluate invariance in rater-mediated assessments and thus, the usefulness of rater-mediated scores for decisions that impact student opportunities.

Research

Findings from this study can inform future research on a variety of topics related to rater-mediated writing assessment. Specifically, findings of differences among indices of rating quality hold significant implications for future studies that use some or all of these quality indices. As a result of differences among rating quality indices, the methods selected by researchers to evaluate the quality of ratings and rater training may affect the interpretation of score differences. Along the same lines, the findings of similarity between the traditional indices of agreement and accuracy and the Rasch-based indices of error and direct accuracy measures should be further investigated with other data sets in order to determine the generalizability of these findings to other writing assessment situations.

In order to understand rater-mediated assessments in terms of a lens model, further research is needed on the cognitive processes of raters during the assessment process. Judgmental processes should be examined for raters who assign high- and low-quality ratings defined by model-data fit and accuracy measures to inform interpretations of scores assigned in rater-mediated assessments. These investigations can provide information about what is and is not reflected in scores as well as factors that may contribute to achievement differences among groups of students on large-scale rater-mediated writing assessments.

Policy and Practice

The implications for findings from this study are, perhaps, most directly related to policy and practice for rater-mediated writing assessment. The correlations revealed that many indicators of rating quality do not provide comparable information for a set of raters. However, the finding that direct measures of accuracy correlate strongly with measures of rater agreement within the traditional perspective and measures of rater error in terms of model-data fit from Rasch measurement theory holds significant implications for improvement of quality-control systems in large-scale rater-mediated assessments. Because limited resources often prevent scoring of each essay by a validity committee or single expert rater in order to directly measure rater accuracy on each student's composition, other indices of rating quality are used as indirect measures of rater accuracy. This study's findings suggest that Rasch-based indices of model-data fit for ratings as well as indices of rater agreement from Facets (Linacre, 2010) provide information about raters that is comparable to direct measures of accuracy. In addition, as many rater-mediated writing assessments transition to online scoring, the possibility to implement continuous accuracy monitoring with interspersed benchmark papers that are rated by experts along with measures of rater agreement and model-data fit is increasing. As demonstrated in the findings from this study, these indices of rating quality can provide useful, comparable information about the quality of ratings that are assigned in rater-mediated writing assessments, and hold great promise for ensuring that all students receive high-quality scores in high-stakes performance assessments.

References

Andrich, D. (2010). *The detection of a structural halo when multiple criteria have the same generic categories for rating.* Paper presented at the international conference on Rasch measurement in Copenhagen, Denmark.

Berkowitz-Jones. (2007). *Examining rater accuracy within the context of a high-stakes writing assessment (Unpublished doctoral dissertation).* Atlanta: Emory University.

Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy*. New York: Springer-Verlag.

Brunswik, E. (1952). *The conceptual framework of psychology.* Chicago: University of Chicago Press.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.) Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance*, *20*, 238-252.

Coffman, W. (1971). Essay examiniations. In R. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 188-201). Washington, DC: American Council on Education.

Cohen (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(37), 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum Associates.

Cole, N. S. (1997). Understanding gender differences and fiar assessment in context. In W. W. Willingham, *Gender and Fair Assessment* (pp. 157-184). Mahwah: Lawrence Erlbaum Associates.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason: Cenage.

Cronbach, L. J. (1947). Test ‘reliability’: Its meaning and determination. *Psychometrika, 12*(1), 1-15.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin, 52*(3), 177-193.

Cronbach, L. (1958). Proposals toward analytic treatment of social perception scores. In R. Tagiuri & L. Petrullo (Eds.), *Person Perception and Interpersonal Behavior* (pp. 353-378). Stanford, CA: Stanford University Press.

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

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*(1), 56-70.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal, *Large-Scale Assessment programs for All Students: Validity, Technical Adequacy, and Implementation* (pp. 261-187). Mahwah: Lawrence Erlbaum Associates.

Engelhard, G. (2007). Differential rater functioning. *Rasch Measurement Transactions, 21*(3) p. 1124.

Engelhard, G. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research & Perspective, 6*, 155-189.

Engelhard, G. (2009). Using item response theory and model data fit to conceptualize differential item functioning for students with disabilities. *Educational and Psychological Measurement, 69*(4), 585-602.

Engelhard, G., Walker, E. V., Gordon, B., & Gabrielson, S. (1994). Writing tasks and gender: influences on writing quality of black and white students. *Journal of Educational Research, 87*, 197-209.

Engelhard, G. & Perkins, A. F. (2011) Person response functions and the definition of units in the social sciences. *Measurement: Interdisciplinary Research & Perspective*, *9,* 40-45.

Georgia Department of Education. (1993). *Georgia grade 8 writing assessment interpretive guide.* Atlanta: Author.

Georgia Department of Education. (1993). *Georgia high school writing test: assessment and instructional guide.* Atlanta: Author.

Gyagenda, I. (1999). *Exploring rater, domain, and gender influences on the assessed quality of student writing using classical and modern measurement theories (Doctoral Dissertation).* Atlanta: Emory University.

Gyagenda, I., & Engelhard, G. (2009). Using classical and modern measurement theories to explore rater, domain, and gender influences on student writing ability. *Journal of Applied Measurement, 10*(3), 225-246.

Hamp-Lyons, L. (1991). The writer's knowledge and our knowledge of the writer. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 51-68). Norwood: Ablex Publishing Corporation.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, 5-16.

Hillocks, G. (2005). At Last: The Focus on Form vs. Content in Teaching Writing. *Research in the Teaching of English, 40*(2), 238-248.

Hogarth, R. M. (1987). *Judgment and choice: The psychology of decision*. (2nd edition). Chichester, England: Wiley.

Hyde, J. (2005). The gender similarities hypothesis. *American Psychologist, 60*(6), 581- 592.

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

Kendall, M. G., & Smith, B. B. (1939). The problem of m rankings. *The Annals of Mathematical Statistics, 10*(3), 275-287.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletein, 87*(1), 72-107.

Linacre, J. M. (2010) Facets Rasch measurement computer program, version 3.67.1. Chicago: Winsteps.com.

Mattern, K., Camara, W., & Kobrin, J. L. (2007). *SAT writing: An overview of research and psychometrics to date.* Office of Research and Analysis. College Board.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

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

Postman, L. & Tolman, E. C. (1959). Brunswik's probabilistic functionalism. In S. Koch (Ed.) *Psychology: A Study of Science*. (pp. 502-564). New York: McGraw-Hill Book Company, Inc.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. (Expanded edition, Chicago: University of Chicago Press, 1980).

Rasch, G. (1968). A mathematical theory of objectivity and its consequences for model construction. In *Report from European Meeting on Statistics, Econometrics and Management Sciences,* Amsterdam, 1968.

Reynolds, D. J., & Gifford, R. (2001). The sounds and sights of intelligence: a lens model channel analysis. *Personality and Social Psychology Bulletin, 27,* 187-200.

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

Schuster, C., & Smith, D. A. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods, 7*(3), 384-395.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*(3), 497-506.

Thurstone, L. L. (1928). Scale construction with weighted observations. *Journal of Educational Psychology, 19*, 441-453.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*(1), 35-51.

Wolfe, E. W. & McVay, A. (2011, April). *Application of Latent Trait Models to Identifying Substantively Interesting Raters.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Chicago: MESA Press.

Zegers, F. E. (1991). Coefficients for Interrater Agreement. *Applied Psychological Measurement, 15*(4), 321-333.

Table 1. Instrument Description

Category	Domain	Definition	Characteristics that Suggest Domain Mastery
Meaning	Organization	The controlling idea is established through examples, illustrations, facts, or details. A sense of order is clear and relevant.	• Response to assigned task • Clearly established controlling idea • Sufficiently relevant supporting ideas • Clearly developed supporting ideas • Clearly discernible order of presentation • Logical transitions and flow of ideas • Sense of completeness
	Style	The writer controls language to establish his or her individuality.	• Effective diction • Varied and effective sentence structure • Tone consistent with topic and purpose • Sense of audience
Mechanics	Conventions	The writer uses the conventions appropriate for standard American written English.	• Appropriate usage • Appropriate mechanics
	Sentence Formation	The writer forms sentences correctly.	• Appropriate end punctuation • Complete sentences or functional fragments • Appropriate coordination and/or subordination

Table 2. Categories of Rating Quality Indices

Category	Traditional Approach	Rasch-based Approach
Agreement	Percent exact rating agreement	N/A
Error and Systematic Bias	Rater means	Rater severity measure Fit statistics for ratings
Accuracy	Percent exact accuracy agreement	Rater accuracy measure Fit statistics for accuracy

Table 3. Summary Statistics for Ratings

	Student (θ)	**Rater (λ)**	**Domain (δ)**
Measure			
M	0.81	0.00	0.00
SD	2.92	0.35	0.69
N	365	20	4
Outfit			
M	1.01	1.02	1.01
SD	0.27	0.18	0.05
Infit			
M	1.00	1.00	1.00
SD	0.22	0.16	0.05
Separation statistic			
Reliability of separation	0.99	0.98	> 0.99
Chi-square (χ^2)	49599.70*	869.50*	2654.20*
(*df*)	(364)	(19)	(3)

* $p < 0.05$

Table 4. Indices of Rater Error and Systematic Bias

	Traditional Approach		**Rasch-based Approach**					
Rater ID	Percent Exact Rating Agreement	Mean rating	Severity Measure (Logits)	SE	Infit MSE	Standardized Infit MSE	Outfit MSE	Standardized Outfit MSE
15	59.30	2.52	0.76	0.05	1.29	7.38	0.75	-5.8
7	55.80	2.58	0.50	0.05	1.02	0.69	1.32	6.21
6	58.40	2.61	0.41	0.05	1.28	7.10	1.10	2.02
10	59.20	2.65	0.26	0.05	1.14	3.70	0.93	-1.47
5	58.30	2.67	0.17	0.05	0.84	-4.78	1.03	0.55
12	58.40	2.67	0.16	0.05	1.00	0.10	1.09	1.78
11	59.10	2.68	0.14	0.05	1.04	1.01	1.04	0.84
21	60.40	2.69	0.10	0.05	1.04	1.19	0.82	-4.04
2	61.00	2.69	0.08	0.05	1.03	0.84	0.81	-4.22
3	55.60	2.70	0.05	0.05	0.89	-3.05	1.24	4.76
19	57.20	2.72	-0.02	0.05	0.84	-4.76	1.08	1.68
4	61.70	2.72	-0.04	0.05	0.84	-6.64	0.80	-4.54
17	54.70	2.73	-0.06	0.05	0.78	-6.36	1.24	4.72
20	61.20	2.73	-0.10	0.05	0.95	-1.30	0.82	-3.86
14	60.40	2.76	-0.19	0.05	1.24	6.21	0.92	-1.77
13	55.80	2.78	-0.29	0.05	1.00	-0.01	1.20	3.74
8	61.60	2.80	-0.36	0.04	1.2	5.29	0.82	-5.21
16	58.60	2.83	-0.47	0.05	0.84	-4.81	1.00	0.09
18	54.30	2.84	-0.52	0.05	1.00	-0.02	1.25	4.44
9	57.00	2.86	-0.59	0.05	0.83	-5.15	1.14	2.48
M	*58.40*	*2.71*	*0.00*	*0.05*	*1.00*	*-0.17*	*1.02*	*0.12*
SD	*2.31*	*0.09*	*0.34*	*0.00*	*0.16*	*4.47*	*0.18*	*3.77*

Note. Raters are ordered by severity measure.

Table 5. Calibration of the Domain Facet

Category	**Domain**	**Measure**	**SE**	**Infit MSE**	**Standardized Infit MSE**	**Outfit MSE**	**Standardized Outfit MSE**	**Mean Rating**
Meaning	Organization	0.68	0.02	0.99	-0.34	1.02	1.01	2.54
	Style	0.32	0.02	1.05	3.39	1.08	3.81	2.64
Mechanics	Conventions	-0.08	0.02	0.93	-4.41	0.95	-2.42	2.74
	Sentence Formation	-0.92	0.02	1.01	0.35	0.99	-0.26	2.94
	M	*0.00*	*0.02*	*1.00*	*-0.25*	*1.01*	*0.54*	*2.72*
	SD	*0.69*	*0.00*	*0.05*	*3.21*	*0.05*	*2.60*	*0.17*

Note. Domains are ordered by Measure (difficulty).

Table 6. Rating Scale Structure

Rating Scale Category	**Label**	**Average Measure**	**Usage (%)**	**Infit MSE**	**Outfit MSE**
1	Inadequate	- 3.97	10	1.05	1.10
2	Poor	- 1.17	29	0.99	1.00
3	Good	1.78	39	0.93	1.00
4	Very Good	4.15	21	1.01	1.00

Note. Categories are ordered by Measure (difficulty).

Table 7. Summary Statistics for Accuracy Ratings

	Raters (β)	**Benchmark Papers (δ)**	**Domains (α)**
Measure			
M	0.00	-0.81	0.00
SD	0.31	0.63	0.06
N	20	365	4
Outfit			
M	1.00	1.00	1.00
SD	0.03	0.09	0.02
Infit			
M	1.00	1.00	1.00
SD	0.01	0.03	0.00
Separation statistic			
Reliability of separation	0.97	0.83	0.84
Chi-square (χ^2)	622.60*	1348.60*	18.60*
(*df*)	(19)	(364)	(3)

* $p < 0.05$

Table 8. Indices of Rater Accuracy

	Traditional Approach	**Rasch-based Approach**					
Rater ID	Percent Exact Accuracy Score Agreement	Accuracy Measure (Logits)	S.E.	Infit MSE	Standardized Infit MSE	Outfit MSE	Standardized Outfit MSE
10	57.90	0.85	0.05	1.01	0.38	1.06	1.41
2	61.30	0.56	0.06	1.00	-0.20	0.98	-0.42
4	58.40	0.38	0.06	1.00	0.04	1.00	-0.04
6	58.30	0.26	0.06	1.02	-0.13	1.04	0.88
11	60.30	0.07	0.06	0.96	-0.20	0.93	-2.11
12	62.00	0.07	0.06	0.99	-0.09	0.99	-0.31
17	58.90	0.05	0.06	1.00	1.05	0.99	-0.32
14	60.50	0.00	0.06	1.00	-0.11	1.00	0.00
8	61.80	-0.04	0.06	0.99	-0.55	1.02	0.49
20	60.40	-0.04	0.06	1.00	-0.28	1.03	1.06
13	62.70	-0.06	0.06	0.99	-1.80	0.96	-1.14
9	62.80	-0.09	0.06	1.00	0.42	0.98	-0.68
16	61.10	-0.10	0.06	1.00	-0.21	1.00	-0.07
3	61.00	-0.14	0.06	1.03	-0.60	1.05	1.71
15	57.50	-0.17	0.06	1.02	-0.76	1.02	0.79
5	60.20	-0.27	0.06	1.00	0.89	0.98	-0.83
18	58.40	-0.29	0.06	1.00	-0.14	1.01	0.52
21	62.10	-0.32	0.06	1.01	0.07	1.01	0.32
7	59.20	-0.35	0.06	0.99	1.57	0.98	-0.80
19	63.00	-0.37	0.06	1.00	-0.10	0.98	-0.85
M	*60.54*	*0.38*	*0.06*	*1.00*	*1.00*	*1.00*	*-0.02*
SD	*1.75*	*1.74*	*0.00*	*0.01*	*0.03*	*0.03*	*0.94*

Note. Raters are ordered by Measure (accuracy).

Table 9

Calibration of Accuracy Ratings within Domains

Category	**Domain**	**Measure** (Difficulty)	**SE**	**Infit MSE**	**Standardized Infit MSE**	**Outfit MSE**	**Standardized Outfit MSE**	**Mean Rating**
Meaning	Style	0.06	0.03	1.00	0.29	1.02	0.99	0.67
	Organization	0.04	0.03	1.00	-0.1	1.01	0.93	0.67
Mechanics	Conventions	-0.03	0.03	1.00	0.27	1.00	-0.01	0.69
	Sentence Formation	-0.07	0.03	0.99	-0.63	0.98	-1.06	0.69
	M	*0.00*	*0.03*	*1.00*	*-0.04*	*0.02*	*0.21*	*0.68*
	SD	*0.06*	*0.00*	*0.01*	*0.43*	*1.00*	*0.96*	*0.01*

Note. Domains are ordered by Measure (difficulty).

Table 10. Correlations among Traditional Indices

Category	**Traditional RQ Indicator**	**(1)**	**(2)**	**(3)**
Agreement	(1) Exact Rater Agreement			
Error	(2) Rater Mean	-0.085		
Accuracy	(3) Exact Accuracy Agreement	0.939*	-0.128	

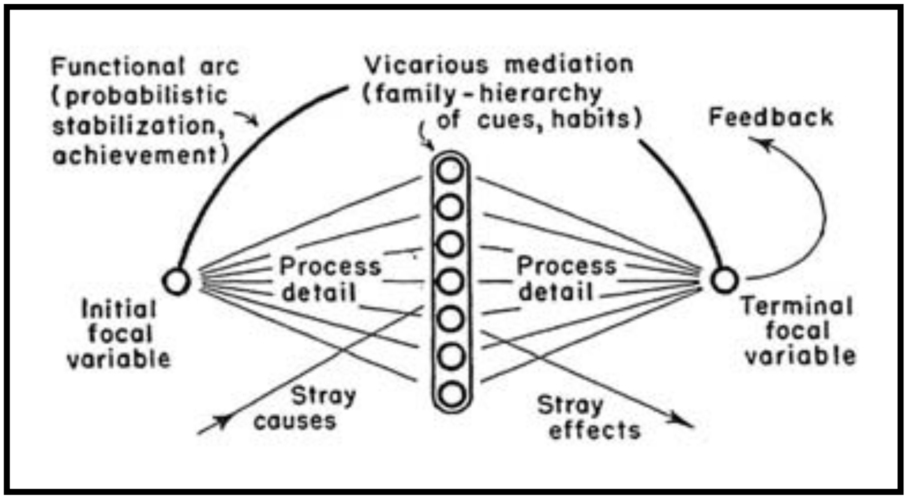
* $p < 0.0033$ (Bonferroni-corrected p value)

Table 11. Correlations among Rasch-based Indices

Category	Rasch-based RQ Indicator	(1)	(2a)	(2b)	(2c)	(2d)	(3)	(4a)	(4b)	(4c)	(4d)
Error	(1) Severity Measure										
	(2) Model-data fit for ratings (2a) Infit MSE	0.430									
	(2b) Standardized Infit MSE	0.420	0.993*								
	(2c) Outfit MSE	-0.157	-0.364	-0.319							
	(2d) Standardized Outfit MSE	-0.131	-0.384	-0.339	0.996*						
Accuracy	(3) Accuracy Measure	-0.172	0.181	0.132	-0.743*	-0.766*					
	(4) Model-data fit for accuracy (4a) Infit MSE	-0.305	-0.188	-0.202	-0.018	0.000	0.178				
	(4b)Standardized Infit MSE	-0.265	-0.004	-0.028	-0.137	-0.121	0.230	0.909*			
	(4c) Outfit MSE	-0.184	0.031	0.034	0.099	0.082	0.269	0.243	0.190		
	(4d)Standardized Outfit MSE	-0.127	-0.025	-0.026	0.115	0.105	0.226	0.176	0.127	0.982*	

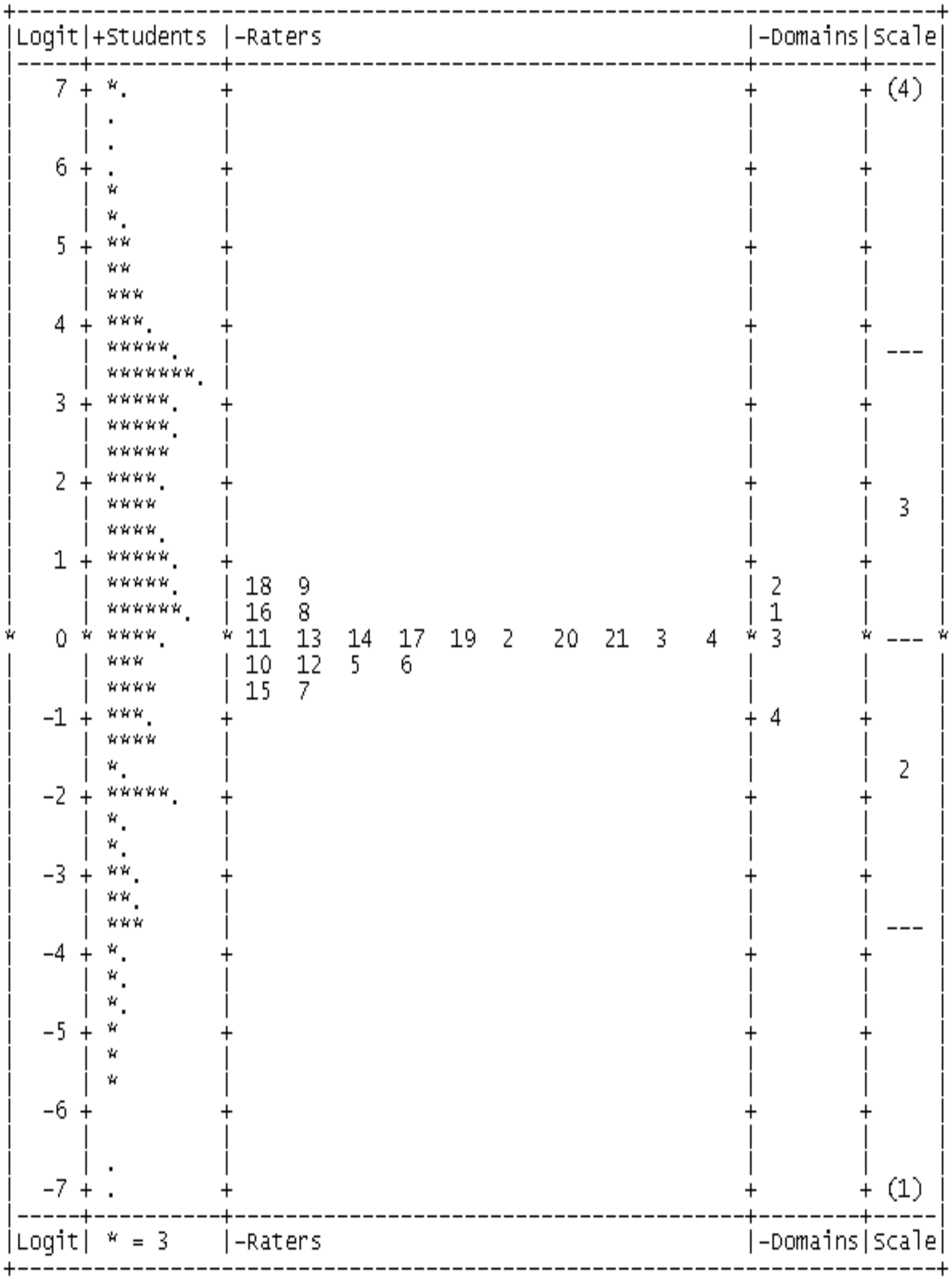
* *p* < .0002 (Bonferroni-corrected *p* value)

Figure 1. Brunswik’s Lens Model for Probabilistic Functionalism



Note. Reproduced from Brunswik, 1952, p. 20

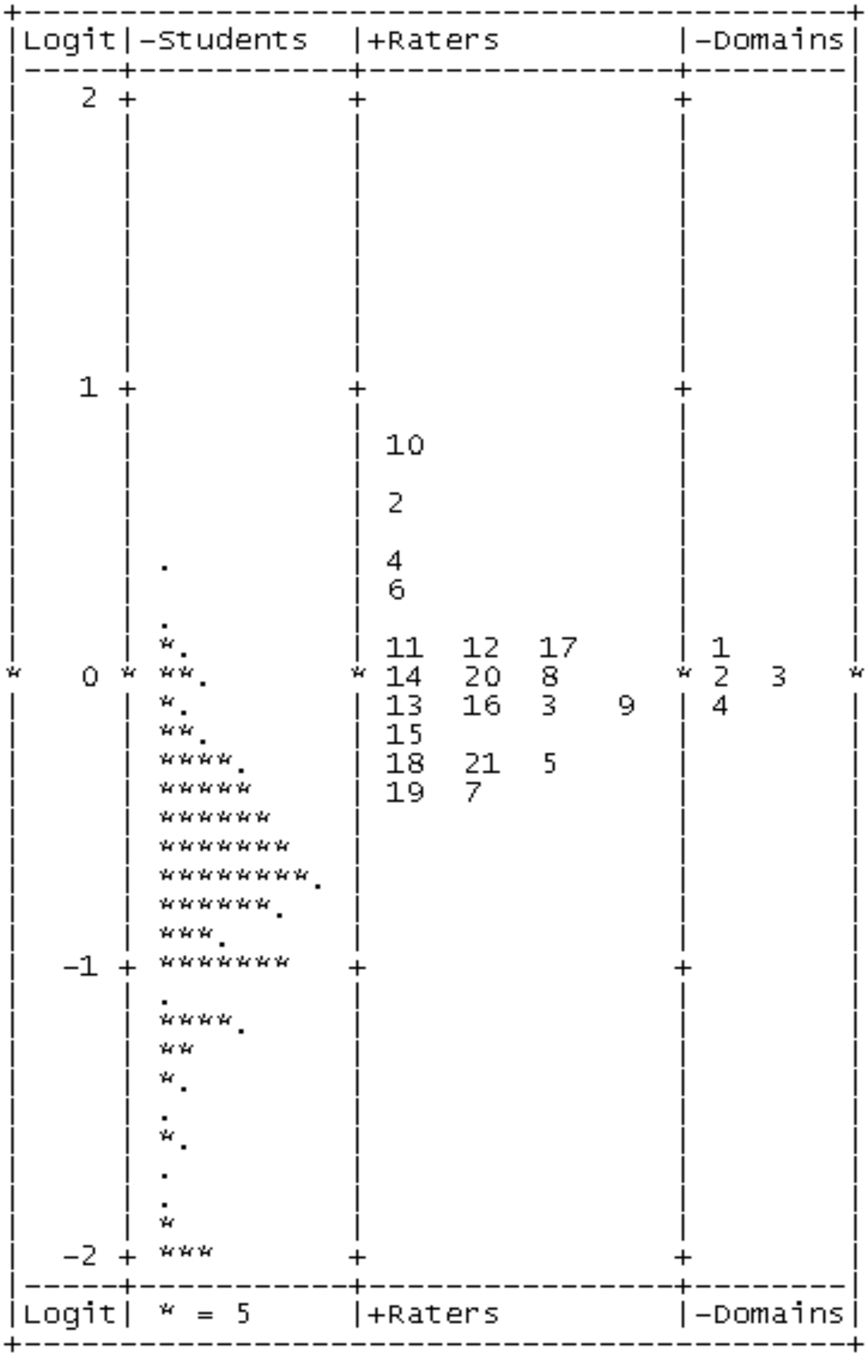
Figure 2. Variable Map for Rating Data



```
+-----------------------------------------------------------------------+
|Logit|+Students |-Raters                                |-Domains|Scale|
|-----+----------+---------------------------------------+--------+-----|
|   7 + *.       +                                       +        + (4) |
|     | .        |                                       |        |     |
|     | .        |                                       |        |     |
|   6 + .        +                                       +        +     |
|     | *        |                                       |        |     |
|     | *.       |                                       |        |     |
|   5 + **       +                                       +        +     |
|     | **       |                                       |        |     |
|     | ***      |                                       |        |     |
|   4 + ***.     +                                       +        +     |
|     | *****.   |                                       |        | --- |
|     | *******. |                                       |        |     |
|   3 + *****.   +                                       +        +     |
|     | *****.   |                                       |        |     |
|     | *****    |                                       |        |     |
|   2 + ****.    +                                       +        +     |
|     | ****     |                                       |        |  3  |
|     | ****.    |                                       |        |     |
|   1 + *****.   +                                       +        +     |
|     | *****.   | 18  9                                 | 2      |     |
|     | ******.  | 16  8                                 | 1      |     |
*   0 * ****.    * 11  13  14  17  19  2   20  21  3   4 * 3      * --- *
|     | ***      | 10  12  5   6                         |        |     |
|     | ****     | 15  7                                 |        |     |
|  -1 + ***.     +                                       + 4      +     |
|     | ****     |                                       |        |     |
|     | *.       |                                       |        |  2  |
|  -2 + *****.   +                                       +        +     |
|     | *.       |                                       |        |     |
|     | *.       |                                       |        |     |
|  -3 + **.      +                                       +        +     |
|     | **.      |                                       |        |     |
|     | ***      |                                       |        | --- |
|  -4 + *.       +                                       +        +     |
|     | *.       |                                       |        |     |
|     | *.       |                                       |        |     |
|  -5 + *        +                                       +        +     |
|     | *        |                                       |        |     |
|     | *        |                                       |        |     |
|  -6 +          +                                       +        +     |
|     |          |                                       |        |     |
|     | .        |                                       |        |     |
|  -7 + .        +                                       +        + (1) |
|-----+----------+---------------------------------------+--------+-----|
|Logit| * = 3    |-Raters                                |-Domains|Scale|
+-----------------------------------------------------------------------+
```

Domains:
1 = Meaning/Style
2 = Meaning/Organization
3 = Mechanics/Conventions
4 = Mechanics/Sentence Formation

Figure 3. Variable Map for Accuracy Data



```
+-------------------------------------------+
|Logit|-Students  |+Raters         |-Domains|
|-----+-----------+----------------+--------|
|   2 +           +                +        |
|     |           |                |        |
|     |           |                |        |
|     |           |                |        |
|     |           |                |        |
|     |           |                |        |
|     |           |                |        |
|     |           |                |        |
|     |           |                |        |
|     |           |                |        |
|   1 +           +                +        |
|     |           |                |        |
|     |           | 10             |        |
|     |           |                |        |
|     |           | 2              |        |
|     |           |                |        |
|     | .         | 4              |        |
|     |           | 6              |        |
|     | .         |                |        |
|     | *.        | 11  12  17     | 1      |
*   0 * **.       * 14  20  8      * 2  3   *
|     | *.        | 13  16  3   9  | 4      |
|     | **.       | 15             |        |
|     | ****.     | 18  21  5      |        |
|     | *****     | 19  7          |        |
|     | ******    |                |        |
|     | *******   |                |        |
|     | ********. |                |        |
|     | ******.   |                |        |
|     | ***.      |                |        |
|  -1 + *******   +                +        |
|     | .         |                |        |
|     | ****.     |                |        |
|     | **        |                |        |
|     | *.        |                |        |
|     | .         |                |        |
|     | *.        |                |        |
|     | .         |                |        |
|     | .         |                |        |
|     | *         |                |        |
|  -2 + ***       +                +        |
|-----+-----------+----------------+--------|
|Logit| * = 5     |+Raters         |-Domains|
+-------------------------------------------+
```

Domains:
1 = Meaning/Style
2 = Meaning/Organization
3 = Mechanics/Conventions
4 = Mechanics/Sentence Formation

Figure 4. Scatter Plots for Traditional Rating Quality Indices

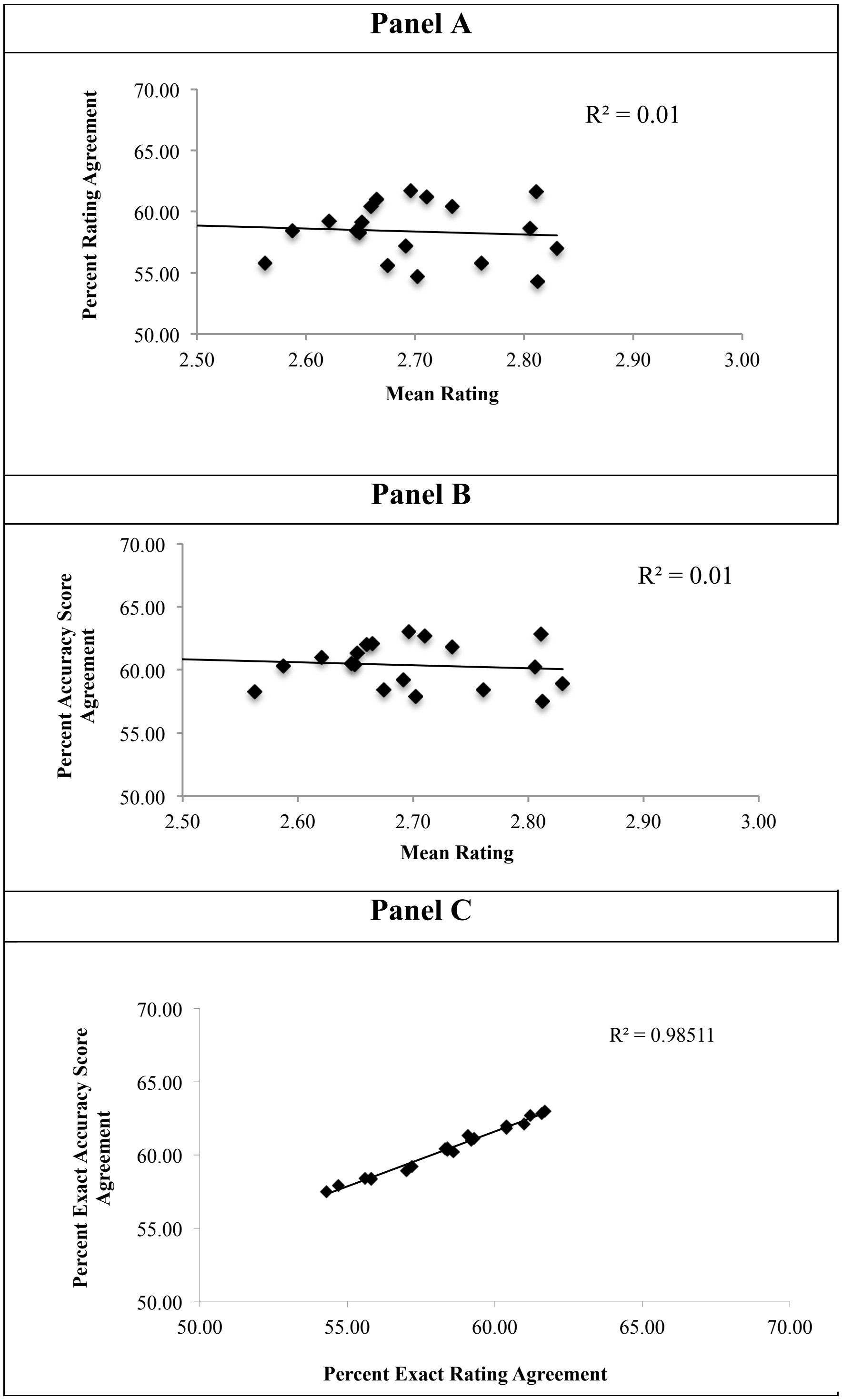
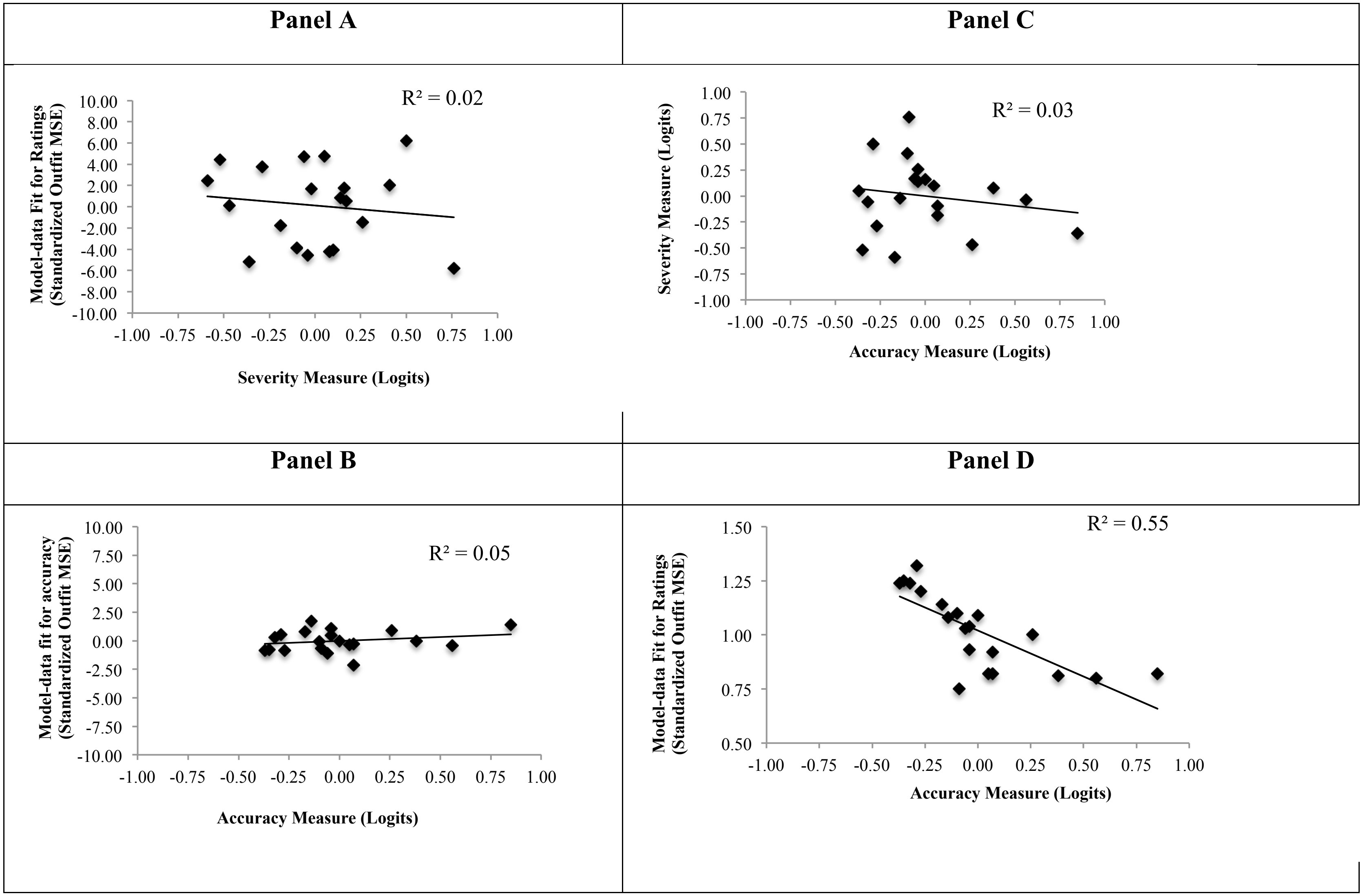
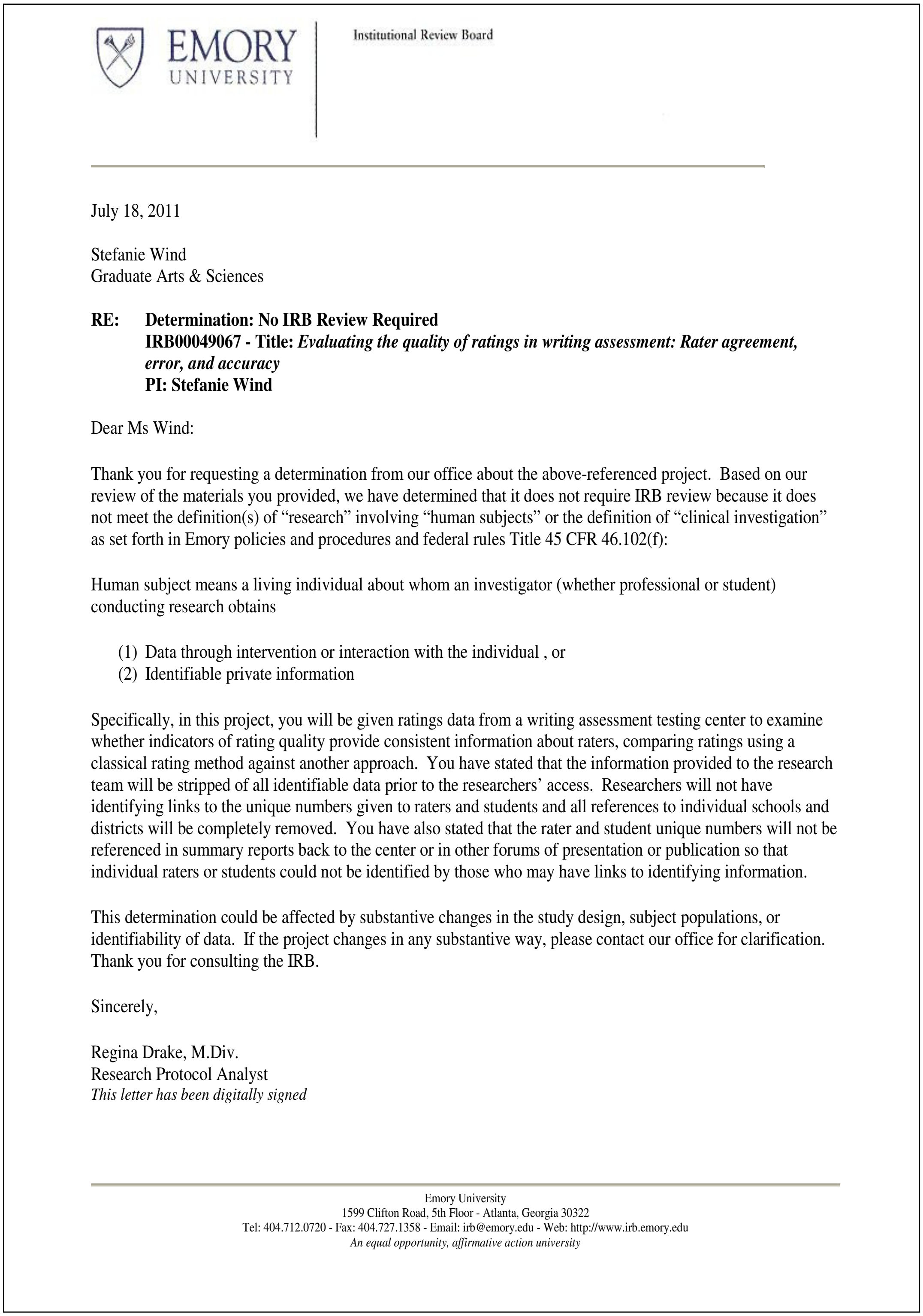


Figure 5. Scatter Plots for Rasch-based Rating Quality Indices



Appendix A

IRB Documentation



EMORY UNIVERSITY

Institutional Review Board

July 18, 2011

Stefanie Wind
Graduate Arts & Sciences

RE: Determination: No IRB Review Required
IRB00049067 - Title: ***Evaluating the quality of ratings in writing assessment: Rater agreement, error, and accuracy***
PI: Stefanie Wind

Dear Ms Wind:

Thank you for requesting a determination from our office about the above-referenced project. Based on our review of the materials you provided, we have determined that it does not require IRB review because it does not meet the definition(s) of "research" involving "human subjects" or the definition of "clinical investigation" as set forth in Emory policies and procedures and federal rules Title 45 CFR 46.102(f):

Human subject means a living individual about whom an investigator (whether professional or student) conducting research obtains

(1) Data through intervention or interaction with the individual , or
(2) Identifiable private information

Specifically, in this project, you will be given ratings data from a writing assessment testing center to examine whether indicators of rating quality provide consistent information about raters, comparing ratings using a classical rating method against another approach. You have stated that the information provided to the research team will be stripped of all identifiable data prior to the researchers' access. Researchers will not have identifying links to the unique numbers given to raters and students and all references to individual schools and districts will be completely removed. You have also stated that the rater and student unique numbers will not be referenced in summary reports back to the center or in other forums of presentation or publication so that individual raters or students could not be identified by those who may have links to identifying information.

This determination could be affected by substantive changes in the study design, subject populations, or identifiability of data. If the project changes in any substantive way, please contact our office for clarification. Thank you for consulting the IRB.

Sincerely,

Regina Drake, M.Div.
Research Protocol Analyst
This letter has been digitally signed

Emory University
1599 Clifton Road, 5th Floor - Atlanta, Georgia 30322
Tel: 404.712.0720 - Fax: 404.727.1358 - Email: irb@emory.edu - Web: http://www.irb.emory.edu
An equal opportunity, affirmative action university