Distribution Agreement

In presenting this thesis or dissertation as partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Stefanie Anne Wind Date

Evaluating Rater-Mediated Assessments

with Rasch Measurement Theory and Mokken Scaling

By

Stefanie A. Wind

Doctor of Philosophy

Division of Educational Studies

George Engelhard Jr., Ph.D.

Advisor

Yuk Fai Cheong, Ph.D.

Committee Member

Robert J. Jensen, Ph.D.

Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.

Dean of the James T. Laney School of Graduate Studies

Date

Evaluating Rater-Mediated Assessments

with Rasch Measurement Theory and Mokken Scaling

By

Stefanie A. Wind
M.A., Emory University, 2012
B.A., University of West Florida, 2009
B.M., University of West Florida, 2009

Advisor: George Engelhard, Jr., Ph.D.

An abstract of a dissertation 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 Doctor of Philosophy
2014

Abstract

Evaluating Rater-Mediated Assessments
with Rasch Measurement Theory and Mokken Scaling

By Stefanie A. Wind

Models based on Rasch Measurement Theory (Rasch, 1960/1980) are frequently used to explore the quality of ratings assigned in large-scale rater-mediated educational assessments (Engelhard, 2013; Wolfe, 2009) because they meet the requirements for invariant measurement. In contrast, the utility of nonparametric models that meet the requirements for invariant measurement for monitoring rating quality is unexplored. Because they are less restrictive, nonparametric models may provide useful information to inform the interpretation and use of rater-assigned scores. The purpose of this study is to describe, illustrate, and extend current indices of rating quality with concepts from Mokken scaling. The major methods used to address the guiding questions for this study include a literature review, illustrative data analyses, and the application of parametric and nonparametric models to data from large-scale rater-mediated assessments. Mokken-based analyses are conducted using the *mokken* package for the *R* statistical software program (van der Ark, 2013; *R* Development Core Team, 2013). Rasch-based analyses are conducted using the Facets program (Linacre, 2010).

Major findings suggest that Mokken scale analysis provide diagnostic information that supplements indices of measurement quality based on Rasch measurement theory. Further, findings suggest that parametric and nonparametric indicators of measurement quality provide related, but slightly different, information about measurement quality in the context of rater-mediated assessments. The diagnostic information provided by the Mokken-based indicators illustrated in this study is especially promising for assessment development, including rater training and the development of scoring rubrics. In response to the increased emphasis on the use of evidence to guide policy and practice in education (Cooper, Levin, & Campbell, 2009; Huff, Steinberg, & Matts, 2010; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002), the use of assessments that require constructed responses (e.g., essays and portfolios) is increasing, such as those included in the next-generation assessments included in the Race to the Top initiative (U.S. Department of Education, 2010). Within the framework of invariant measurement, this study proposes and applies a coherent set of indicators of rating quality based on measurement models with useful properties that can be used in practice to inform the development, interpretation, and use of rater-mediated assessments.

Evaluating Rater-Mediated Assessment
with Rasch Measurement Theory and Mokken Scaling

Stefanie A. Wind
M.A., Emory University, 2012
B.A., University of West Florida, 2009
B.M., University of West Florida, 2009

Committee Members:
Dr. George Engelhard, Jr., Chair
Dr. Yuk Fai Cheong
Dr. Robert Jensen

A dissertation 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 Doctor of Philosophy
2014

Acknowledgements

“I thank my God in all my remembrance of you,
always in every prayer of mine …” Philippians 1:3

To the children, teachers, and raters who contributed the data that are explored in this dissertation: Thank you for teaching me.

To my family, especially Mom, Dad, Eric, Nik, and Emily: Thanks. My appreciation for you is immeasurable.

To Dr. Engelhard: Thank you for being my advisor, mentor, teacher, musical collaborator, travel companion, and friend. Your generosity, kindness, and wisdom have helped to shape me as a thinker, writer, and person. I look forward our future measurement adventures!

To Grace Lutheran Church: When I moved to Atlanta, I expected to earn a Master’s degree, and hopefully a PhD, but I did not expect to find a family. Thank you for your friendship, encouragement, and teaching (and the HDFs).

To the readers of this dissertation: Thank you for joining me in this psychometric journey, and for recognizing the limitations of this dissertation as important steps towards a more-complete understanding of rater-mediated assessments. There is much work to be done.

Table of Contents

List of Tables, Figures, and Appendices

List of Abbreviations and Acronyms used Throughout the Study

- *AISP*: Automated item selection procedure
- *ANOVA*: Analysis of variance
- *CRF*: Category response function
- *CTT*: Classical test theory
- *D study*: Decision study
- *DM model*: Double Monotonicity model
- *DM-R model*: Double Monotonicity for Ratings model
- *G study*: Generalizability study
- *G theory*: Generalizability theory
- *ICC*: Intraclass correlation coefficient
- *IIO*: Invariant item ordering
- *IRF*: Item response function
- *IRT*: Item response theory
- *MFR model*: Many-facet Rasch model
- *MH model*: Monotone Homogeneity model
- *MH-R model*: Monotone Homogeneity for Ratings model
- *MIIO*: Manifest invariant item ordering
- *MIRO*: Manifest invariant rater ordering
- *MS statistic*: Molenaar-Sijtsma statistic
- *MSE*: Mean square error
- *NIRT*: Nonparametric item response theory

- *OCF*: Operating characteristic function
- *PC model*: Partial credit formulation of the Rasch model
- *PRF*: Person response function
- *RRF*: Rater response function
- *RS model*: Rating scale formulation of the Rasch model
- *SDM model*: Strong double monotonicity model
- *SLM*: Simple logistic model
- *SOL*: Stochastic ordering on the latent variable
- *2PL model*: Two-parameter logistic model
- *3PL model*: Three-parameter logistic model

Chapter One: Introduction

In response to the increased emphasis on evidence to guide policy and practice in education (Cooper, Levin, & Campbell, 2009; Huff, Steinberg, & Matts, 2010; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002), the use of assessments that require constructed responses is increasing (Lane & Stone, 2006). Many current large-scale assessments that are implemented worldwide extend beyond selected-response items and require students to perform a task such as composing an essay or creating a portfolio. Some salient examples include the Program for International Student Assessment (PISA; OECD, 2012), the Test of English as a Foreign Language (TOEFL; Educational Testing Service, 2010), the International Association for the Evaluation of Educational Achievement's (IEA) studies in mathematics (TIMSS; Trends in International Mathematics and Science Study; Mullis, Martin, Foy, & Arora, 2012) and reading (PIRLS; Progress in International Reading Literacy Study; Mullis, Martin, Foy, & Drucker, 2012), and the "next-generation assessments" that are part of the Race to the Top initiative in the United States (U.S. Department of Education, 2010). These assessments provide opportunities for students to construct a response that is judged by a rater according to a rating scale that is designed to represent a construct. A variety of terms are frequently used to describe rater-mediated assessments that emphasize different aspects of these measurement procedures. For example, the terms *constructed-response assessment* (Bennett, 1993) and *performance assessment* (Lane & Stone, 2006) emphasize differences between these assessments and their multiple-choice or selected-response counterparts. On the other hand, the terms *authentic assessment* (Wiggins, 1989) and *direct assessment* (Huot, 1990) focus on the emulation of the context or "real-world" conditions in which the knowledge or skills being assessed are typically applied. It is important to recognize that ratings are not a direct representation of a student in terms of a construct—rather, scores on constructed-response tasks are mediated through human raters who exist within complex ecological contexts. For this reason, this study uses the term *rater-mediated assessments* to describe assessments that include constructed-response tasks that are scored by human raters using a rating scale.

In general, the use of rater-mediated assessments reflects a view that a rater's judgment of a response provides information beyond what could be provided by a more "objective" measure, such as a set of multiple-choice items. Based on the concept of pedagogical *washback* (Hamp-Lyons, 2002; Messick, 1996), proponents of rater-mediated assessments often view these assessments as tools of educational reform that have the potential to encourage more meaningful pedagogical practices than selected-response assessments (Lane and Stone, 2006). However, the usefulness of rater-assigned scores for informing educational decisions depends on the degree to which rater-mediated assessment systems demonstrate useful psychometric properties, including validity, reliability, and fairness. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, in preparation) highlight the fundamental nature of validity for the development and use of educational and psychological measures. As stated in the *Standards*, "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (p. 1). Current research on the concept of validity stresses the use of test scores (Kane, 1992, 2001), and the development of evidence-centered designs to support validity arguments (Huff, Steinberg, & Matts, 2010; Mislevy et al., 2002). However, these aspects of validity tell only part of the story. As pointed out by Messick (1995), validity studies should also address *score meaning***,** and explicitly recognize that score meaning is a function of persons and items, as well as contextual aspects of the assessment. In his words:

> Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions, but also of the persons responding as well as the context of the assessment. In particular, what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails (Cronbach, 1971). The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question (p. 741).

In order to develop a community of meaning and use around a rater-mediated assessment, it is necessary to consider the unique challenges that arise when raters are introduced to the measurement system. Because rater-mediated assessments are complex systems that involve the combination of a variety of facets, it is essential that methods used to evaluate the quality of these assessments account for each of these components. This study examines and extends methods for evaluating the quality of rater-mediated measurement in the context of high-stakes rater-mediated educational assessments.

Theoretical Framework

In order to explore the usefulness of any measurement system, it is necessary to establish an overarching theoretical framework within which aspects of the system can be considered and evaluated. Figure 1 illustrates the three major components that define the theoretical framework for this study: 1) a theory of human judgment, 2) a theory of measurement, and 3) evidence of rating quality. This chapter introduces the theoretical framework and considers theories of human judgment and theories of measurement as they apply to the study. Evidence of rating quality is the major focus of this study, and it is explored further in Chapters Two through Five.

Theories of Human Judgment

The first component of the theoretical framework for this study is a theory of human judgment. Essentially, a theory of human judgment provides a framework for describing relationships among variables that influence human judgment and decision-making activities, such as the judgmental procedures that occur during rater-mediated assessments when raters score student responses.

The scientific study of human judgment has origins in early 20th-century work on psychophysics. Specifically, Fechner's (1860) *Elemente der Psychophysik* is said to have marked the beginning of the systematic use of mental measurement. Expanding upon Weber's (1846/1912) concept of *just noticeable differences* (JNDs) that describe thresholds in human sensations of differences among physical stimuli, Fechner's treatise establishes a logarithmic relation between sensation and changes in the strength of physical stimuli with an external "known" value. These early psychophysicists focused on developing a general mathematical law to describe the functional relationship between human sensations and physical stimuli, such as weight, brightness, or the intensity of sound. The primary assumptions underlying this early work in psychophysics—that human sensations differ within and across persons, and that these sensations are subject to error—led to the development of a variety of theories and procedures to explore potential causes for distorted judgment of both physical stimuli and non-physical stimuli, such as beliefs, attitudes, or achievement (e.g., Thurstone, 1928).

Lens Models

The notion that human judgment varies as a result of mediating variables remains a central component in research on human judgment within the field of cognitive psychology. Since around the 1950s, lens models have been widely used in investigations that consider the role of ecological context on human judgment and decision-making (Karelaia & Hogarth, 2008). Specifically, Brunswik's (1952) lens model has been applied to research on human judgment within a variety of contexts, including clinical diagnosis (Hammond, 1955) and education (Cooksey, Freebody, & Bennett, 1990). The theoretical framework for this study uses Brunswik's (1952) lens model to guide the interpretation of rater judgments within the context of rater-mediated educational assessments.

Brunswik's (1952) Lens Model. Brunswik's (1952) lens model for probabilistic functionalism is depicted in Figure 2. This lens model and its extensions focus on the influence of the ecological context within which a variable is observed on the accuracy of observations of that variable. Essentially, the lens model provides a visual representation of the impact of various cues, or mediating variables, on perceptions of a variable. Describing the tendency in psychological research to ignore the ecological context within which observations are made, Brunswik (1952) notes that complete understanding of behavior requires examination of the conditions and supports for the behavior that are provided within a particular context. In his words, "psychology has forgotten that it is a science of organism-environment relationships, and has become a science of the organism" (Brunswik, 1957, p. 6). The lens model provides a method for describing the impact of a variety of mediating variables through which events are perceived. In Brunswik's (1952) words:

> 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. (p. 91)

Concern with the accuracy of human judgment and decision-making contributed to the widespread application of Brunswik's (1952) lens model in social science research (Goldstein, 2004). Of particular importance is Hammond's (1955) extension of the lens model to the area of clinical diagnosis that eventually led to the development of Social Judgment Theory (Cooksey, 1996a, 1996b; Hammond, Stewart, Brehmer, & Steinmann, 1975). As stated by Cooksey (1996a), Social Judgment Theory provides "a methodology and a perspective for understanding human judgment as it was exercised within a particular ecological context" (p. 141). Hammond and Joyce (1975) describe Social Judgment Theory in detail. Goldstein (2004) suggests the work of several key authors for thorough descriptions of its applications and extensions[1].

Lens Models for Ratings

The study of human judgments in social science research is typically based on ratings collected using surveys or evaluative judgments (Landy & Farr, 1980). Literature that examines rating quality often highlights raters and rating scales, as a type of lens, or

[1] Goldstein (2004, p. 38) lists the following authors as key sources for the application of Brunswik's (1952) approach to judgment: Brehmer & Joyce, 1988; Doherty, 1996; Hammond, 1996, 2000; Hammond & Joyce, 1975; Hammond & Stewart, 2001; Hammond & Wascoe, 1980; Juslin & Montgomery, 1999; Rappoport & Summers, 1973.

filter, through which a student's response is viewed. For example, in their literature review on ratings in performance assessment, Landy and Farr (1980) concluded that rater-assigned scores 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" to describe a person in terms of a construct (p. 100). Further, Engelhard (2013) presents a conceptual framework for examining rating quality based on lens models. Considering a rater-mediated assessment in terms of Brunswik's (1952) model, he notes that mediating cues may interfere in the assessment situation and distort rating quality. These cues may include characteristics of raters (e.g., rater severity), the assessment (e.g., domains), and the scoring system (e.g., the rating scale). Within a particular ecological context, he notes that "the intervening variables define a 'lens' that raters use to focus their judgments and inferences abut person locations on the latent variable" (p. 194). Eckes (2011) describes a similar lens model view of rater-mediated assessments in which he describes cues that are directly related to a construct as *proximal*, and cues that are construct-irrelevant as *distal*.

Figure 3 extends the lens model to the context of a rater-mediated assessment; this lens model is an adaptation of Engelhard's (2013) lens model for rater-mediated assessments. In the language of Brunswik (1952), the initial focal variable (θ_P) is a student's true location on a latent variable, and the rater's judgment about the student's performance (θ_R) is the terminal focal variable. Several things are important to observe in the extension of the lens model to rater-mediated assessment. First, the judged location (θ_R) is informed by cues that include aspects of the assessment system. The cues may include domains on an analytic rubric, student benchmark performances that represent levels of achievement specific to a particular assessment, and the rating scale that corresponds to the rubric for a particular assessment. Second, the types and role of these cues are context-specific, and each rater-mediated assessment system must be considered in terms of its unique ecological context. Finally, rating quality can be considered from the perspective of a lens model using the match between θ_R and θ_P to describe the proximity of rater's interpretation of a student's performance to their true location on the construct of interest. Indicators of rating quality are the focus of this dissertation.

Theories of Measurement

Although information about the context in which ratings are collected informs the interpretation and use of rater-assigned scores, additional tools are needed to clarify the relationships among mediating variables that define the context of a rater-mediated measurement system. Specifically, the interpretation and use of results from a rater-mediated assessment requires a theory in which to bring together a potentially disparate set of variables in a systematic way. As such, the second component of the theoretical framework for this study is a measurement theory that can be used to describe various aspects of rater-mediated assessment. Engelhard (2013) draws upon the work of Messick (1983) and Lazarsfeld (1966) to describe characteristics of measurement theories that can be viewed within the organizing framework of research traditions (Laudan, 1977). Essentially, measurement theories are a combination of a conceptual framework and statistical machinery that provides a system for drawing inferences from scores. Messick's (1983) definition of a measurement theory is as follows:

> Theories of measurement broadly conceived may be viewed as loosely integrated conceptual frameworks within which are embedded rigorously formulated statistical models of estimation and inference about the properties of measurements and scores. (p. 498, cited in Engelhard, 2013, p. 79)

Lazarsfeld (1966) describes the role of measurement theories to guide the use and interpretation of inferences based on statistical models:

> Problems of concept formation, of meaning, and of measurement necessarily fuse into each other … measurement, classification and concept formation in the behavioral sciences exhibit special difficulties. They can be met by a variety of procedures, and only a careful analysis of the procedure and its relation to alternative solutions can clarify the problem itself, which the procedure attempts to solve. (p. 144, cited in Engelhard, 2013, p. 80)

Synthesizing Messick (1983) and Lazarsfeld (1966), Engelhard (2013) summarizes the importance of measurement theories. In his words, the role of measurement theories is to:

- define the aspects of quantification that are defined as problematic,
- determine the statistical models and appropriate methods used to solve these problems,
- determine the impact of our research in the social, behavioral, and health sciences,
- frame the substantive conclusions and inferences that we draw, and ultimately
- delineate and limit the policies and practices derived from our research work in the social, behavioral, and health sciences. (p. 80).

Measurement theories are situated within research traditions. According to Engelhard (2013), research traditions are similar to the concept of paradigms (Kuhn, 1970), scientific research programs (Lakatos, 1978), and disciplines (Cronbach, 1957, 1975). Research traditions are used to identify measurement problems, define methods for addressing these problems, and examine the impact of the problems and solutions on social science research. This section presents an overview of theories of measurement that are applied to rater-mediated assessments within two major research traditions: 1) the test score tradition, and 2) the scaling tradition.

Test Score and Scaling Traditions

Two major research traditions are used to organize theories of measurement in this chapter: 1) the test score tradition, and 2) the scaling tradition (Engelhard, 2008, 2013). With origins in the work of Spearman (1904), the test score tradition focuses on identifying and decomposing sources of error in order to understand the relationship between observed scores and true scores. Measurement theories within the test score tradition that have been applied to raters include Generalizability Theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), factor analysis (Harman, 1976), and structural equation modeling (Jorsekog, 2007) as methods for examining the impact of a variety of sources of measurement error on ratings (Clauser, Clyman, & Swanson, 1999; Harik, et al., 2009; Schoonen, 2005). On the other hand, the scaling tradition has roots in the work of Thorndike during the early 1900s that focuses on creating variable maps to represent a visual display, or “ruler,” on which to operationally define a variable (Thorndike, 1904). Measurement models within the scaling tradition are used to locate persons, items, and other aspects of measurement systems on a common scale that represents a latent variable. Within the scaling tradition, models based on Item Response Theory (IRT) have been applied to rater-mediated assessments in order to calibrate raters and students on a single scale that represents an underlying construct. Essentially, IRT models describe the relationship between a person's location on the latent variable and the probability for a response. In particular, IRT models based on Rasch Measurement Theory have been applied to rater-mediated assessments because they allow for the simultaneous placement of raters, students, and other aspects of a rater-mediated assessment system, such as domains and prompts, on a common scale (Engelhard, 2002; Wolfe, 2009). Because IRT allows for the calibration of items—and, by extension, raters—and students on a single scale, it is possible to obtain measures of persons that are independent of raters, and calibrations of raters that are independent of persons. This is the fundamental property of invariant measurement.

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 psychophysics and psychometrics, recognized the need for objectivity through invariant measures. Calling for invariance of scales across groups of persons, 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). In essence, invariant measurement is based on the idea that measures of phenomena of interest must not be impacted by irrelevant characteristics of the process used to collect those measures. Within the context of educational measurement, phenomena of interest are constructs, and the processes used to collect measures are typically assessment items or tasks. If invariant measurement is not achieved within a measurement system, persons will appear to possess “more” of the trait being measured on tests that are composed of easier items, and persons will appear to possess “less” of the trait being measured on tests that are composed of harder items.

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). Wright and Stone (1979) describe the concept of invariant measurement in terms of requirements for the measurement of persons and the calibration of items. In their words:

> The calibration of test-item difficulty must be independent of the particular persons used for the calibration. The measurement of person ability must be independent of the particular test items used for measuring. When we compare one item with another in order to calibrate a test it should not matter whose responses to these items we use for the comparison…. When we expose persons to a selection of items in order to measure their ability, it should not matter which selection of items we use or which items they complete. (p. xii)

Engelhard and Perkins (2011) expand the conditions for invariant measurement to a set of 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, 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 adherence to these requirements is observed empirically, person and item estimates can be described on a linear scale that represents the variable of interest, and their locations can be compared. In a sense, the fifth requirement embodies the first four: Unidimensionality implies that a single line representing a single dimension of the trait of interest is a useful and apt description of what is observed when a set of persons responds to a set of items.

Ideal-Type Models

This set of requirements can be examined through the use of measurement models that meet the requirements for invariant measurement. As will be seen in Chapter Three and Chapter Four, models exist within both Rasch measurement theory (Rasch, 1960/1980) and Mokken scale analysis (Mokken, 1971) that meet these requirements related to person ordering, item ordering, and unidimensionality. Engelhard (2008) describes measurement models that adhere to the requirements for invariant measurement as *ideal-type models.* In contrast to an approach in which data are reproduced using a variety of models that are tailored to match the idiosyncrasies of a particular dataset, ideal-type models emphasize an ideal structure for measurement that focuses on fitting data to a model that is guided by an underlying set of useful measurement properties (Engelhard, 2013). Ideal-type measurement models specify the rules for measurement a priori and hold data “accountable” to meeting the pre-specified requirements before the measurement model is used to produce meaningful measures for items or people. When acceptable fit to an ideal-type model is observed in data, invariant measurement produces desirable qualities of the measurement scale that can provide credence to results found in later statistical analyses. Without proper theoretical grounds for measurement as required by this property, analysis of results from the measurement instrument can lead to faulty inferences about persons or groups (Wright & Stone, 1979).

Rater-Invariant Measurement

When raters are introduced into the measurement system, it is possible to examine whether or not data meet the requirements for invariance. In this section, the duality between items and persons that characterizes the requirements for invariant measurement is extended to the concept of *rater-invariant measurement,* which emphasizes duality between rater-invariant measurement of persons and person-invariant calibration of raters. Four requirements are stated for rater-invariant measurement:

Rater-Mediated Person Measurement

1. The measurement of persons must be independent of the particular raters who happen to be used for the measuring: *Rater-invariant measurement of persons*.

In the context of rater-mediated measurement, it is easy to imagine that some raters are more severe than others, while others are more lenient. As a result, students whose work is rated by a lenient rater may have an advantage over those students whose work is rated by a severe rater. The goal of rater-invariant measurement of persons is that the rating(s) that are assigned to a particular student can be used to measure that student in terms of the latent variable, and are not dependent upon the "luck of the rater draw."

Rater-Mediated Domain Calibration

2. The calibration of the domains must be independent of the particular raters used for calibration: *Rater-invariant calibration of domains.*

The second requirement for rater-invariant measurement is related to the use of analytic rating procedures, or rating scales that require raters to assign separate ratings related to distinct aspects of a student's performance, such as mechanics, content, and organization in writing. In order to achieve rater-invariant measurement, it is necessary that the meaning of an analytic rating scale remains consistent across a group of students. Figure 4 illustrates this requirement through a graphical display of rater-*invariant* calibration of domains and rater-*variant* calibration of domains. Specifically, the display depicts ratings in three domains: Mechanics (M), Content (C), and Organization (O). Panel A represents a rater whose interpretation of the rating scale is consistent across student achievement levels, while Panel B represents a rater whose interpretation of the rating scale varies across student achievement levels. In Panel A (rater-invariant measurement), the ordering of the three domains is invariant with the mechanics (M) domain judged easiest and organization (O) domain judged as hardest across the latent variable of writing proficiency. In Panel B (rater-variant measurement), the meaning of domain difficulty varies as a function of person location on the latent variable of writing proficiency. The invariant rater interprets the domains in a comparable way over subgroups with domains ordered as M < C < O, while the domain difficulties are not comparable over subgroups for the variant rater. The variant rater rates organization (O) as the easiest domain for persons with low writing proficiency, while organization (O) is rated as the hardest domain for persons with high writing proficiency. This idea is described further in Wind and Engelhard (2011).

Rater-Mediated Rating Categories

3. The structure of the rating categories must be independent of the particular raters used for calibration: *Rater-invariant calibration of rating scales*.

The third requirement is related to raters' use of rating scale categories when assigning polytomous ratings. Rating scales with multiple categories allow raters to distinguish among students at various levels of a construct. As a result, rating scales can be considered a method for partitioning a latent variable into adjacent intervals that describe substantively meaningful differences among students in terms of a construct (Andrich, de Jong, & Sheridan, 1997). In order to achieve rater-invariant measurement, it is necessary that rating scale categories have a comparable meaning across a group of raters. Figure 5 illustrates this requirement for two raters who do not meet the requirement of rater-invariant calibration of rating scales. The star shape is used to represent a student's location on the latent variable. As can be seen in the figure, the same location on the latent variable results in two different ratings when the structure of the rating categories is not invariant over individual raters. If Rater A rates the student, their location on the construct is judged as a "minimal" performance. However, if Rater B rates the student, the same location on the construct is judged as "good."

Rater-Mediated Variable Map

4. Persons, raters, domains, and rating categories must be simultaneously located on a single underlying latent variable: *Variable map.*

Fourth, rater-invariant measurement requires that persons, raters, domains, and rating scale categories be measured on a single scale that represents the latent variable. When these aspects of a rater-mediated measurement system are located on the same scale, it is possible to create a visual display to represent the operational definition of the latent variable; this visual display is known as a *variable map.* Figure 6 is an illustrative variable map for a writing assessment. The first two columns represent the locations of three students (θ_A, θ_B, θ_C) on the latent variable, which is described using measures on a logit (i.e., log-odds) scale. Next, the domains, benchmarks, and rating scale categories are mapped onto the logit scale. These three facets of the rater-mediated assessment system are labeled as "cues" in order to link the requirements of rater-invariant measurement to the lens model for rater-mediated assessment that was introduced earlier. Finally, Rater λ is calibrated on the logit scale in terms of their overall severity when scoring these three essays. This variable map captures the essence of rater-invariant measurement, and it provides an operational definition of a latent variable within the context of a rater-mediated measurement system.

Ideal-Type Models for Raters

Adherence to the requirements for rater-invariant measurement provides evidence that raters are assigning scores with useful measurement properties. Previous research has applied models based on Rasch Measurement Theory (Rasch, 1960/1980) to the context of rater-mediated assessment. Rasch models are parametric IRT models[2] that meet the

[2] The distinction between parametric and nonparametric models in the context of Item Response Theory is explored in Chapter Four.

requirements of invariant measurement. In contrast, Mokken (1971) proposed a nonparametric IRT model that meets these requirements without the restriction that a population of persons or items must match a particular distribution shape. This study extends Mokken's (1971) theory and procedure for scale analysis to the context of rater-mediated assessments.

Statement of the Problem

The increased use of performance assessments to inform high-stakes educational decisions establishes a need for rating quality indices that can be used in practice to inform the interpretation and use of rater-assigned scores. When applied to data from rater-mediated assessments, models that meet the requirements for invariant measurement provide a method for evaluating the degree to which raters assign scores with useful measurement properties. Although rating quality has been examined using indicators based on parametric IRT models, the utility of nonparametric IRT models for monitoring rating quality is unexplored. Nonparametric techniques can be used for analyses of item response data with less room for improper interpretations and use that may result from violated requirements than their parametric counterparts. The application of nonparametric IRT models to rater-mediated assessments is promising in light of the fact that "if an IRT model is used for constructing a test, and the measurement of respondents on an *ordinal scale* is sufficient for the application envisaged, parametric models might be unduly restrictive for this purpose," and the fact that desirable measurement properties such as invariant ordering of persons and items can be obtained under nonparametric IRT models (Sijtsma & Molenaar, 2002, p. 15).

Purpose of the Study

The purpose of this study is to describe and extend current indices of rating quality with concepts from Mokken scaling. Specifically, indices of rating quality based on test score and scaling traditions are reviewed and considered alongside a new set of rating quality indices based on nonparametric IRT models from Mokken. Applications of these rating quality indices to data from large-scale rater-mediated assessments are used to consider the usefulness and implications of nonparametric IRT models for raters.

Research Questions

This study is guided by five overarching research questions:

1. What are the major underlying measurement issues related to rating quality?
2. How have these measurement issues been traditionally addressed in previous research?
3. How has Rasch measurement theory been used to examine the quality of ratings?
4. How can Mokken scaling be used to examine the quality of ratings?
5. What is the relationship between Mokken- and Rasch-based indices of rating quality?

Definitions

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

Category response function (CRF): The CRF describes the functional relationship between the probability of earning a rating in Category *k* or higher and the difficulty of a rating scale category. The CRF may be defined using cumulative probabilities, as in the case of Mokken's nonparametric models, or using conditional probabilities, as in the Partial-Credit formulation of the Rasch model.

Conditional independence: Responses to an item are not influenced by responses to any other item, after controlling for the latent variable.

Conditional rater independence: The rating assigned to a student is not influenced by ratings assigned by other raters.

Domain: An aspect of performance that is believed to be conceptually distinct from other aspects of performance, such as meaning vs. mechanics in writing. Domains are often scored separately using analytic rubrics.

Expert raters and validity committees: Individuals or groups of raters whose expertise is considered sufficient for the assignment of scores that reflect “true” or “accurate” measures of a student’s achievement. Scores assigned by expert raters and validity committees are used as criteria for the evaluation of the quality of scores assigned by operational raters (defined below).

Facets: Explanatory variables, such as raters, tasks, and assessment occasions, that are incorporated into Many-Facet Rasch (MFR) models (Linacre, 1989/1994).

Facets computer program: 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 the parametric analyses in this study.

Guttman error: A response pattern involving two items where a positively keyed response ($X = 1$) is observed for the more difficult item and a negatively keyed response ($X = 0$) is observed for the easier item; for example, if Item i and Item j are ordered by difficulty such that Item j is easier than Item i ($i < j$), a score pattern of ($i, j = 0,1$) is a Guttman error. Counting and weighting Guttman errors is the basis for calculating scalability (*H*) coefficients within the framework of Mokken scale analysis.

Guttman scale: A set of item responses in which a person's total score can be used to reproduce the exact item responses in the data matrix.

Ideal-type models: Models that meet the requirements for invariant measurement; these models are guided by an underlying set of useful measurement properties (defined below).

Item response function (IRF): The functional relationship between the probability of providing a correct or positive response to an item and the difficulty of an item. The IRF is also referred to as an Item Characteristic Curve (ICC).

Interrater reliability: A measure of the equivalence in the rank-ordering of performances among a group of raters.

Intrarater reliability: A measure of the consistency of individual raters within their own ratings.

Model-data fit: The match between observable properties in data and the assumptions or requirements of a model. Evidence for adequate model-data fit suggests that a model is an appropriate summary of a dataset.

Mokken scale: A set of items (or, in the case of this study, a set of raters) that can be ordered such that the overall scalability coefficient (*H*) is larger than a specified critical value (usually $H \geq 0.30$).

mokken package: A statistical software package for the *R* computer program (R Development Core Team, 2013) that is used to implement techniques based on Mokken scale analysis (Mokken, 1971) in this study. Version 2.7.5 of *mokken* (van der Ark, 2013) is used for the nonparametric analyses in this study.

Monotonicity: A monotonic relationship exists when an increase in the latent variable corresponds to an increase in raw (i.e., observed) scores.

Nonparametric item response theory models: A class of item response models whose functional form assumptions do not require adherence to a specified algebraic form.

Operating characteristic function (OCF): The functional relationship between the probability of a correct response and the logit scale that represents the latent variable. OCFs may be specified for persons, items, or raters; the definitions of these functions vary depending on how the *x*-axis is operationalized (Samejima, 1983).

Parametric item response theory models: A class of item response models whose functional form assumptions (requirements) require adherence to a specified algebraic form (usually the normal or logistic function).

Rater accuracy: The degree to which raters assign scores equivalent to “true” or “known” scores. Often, accuracy is estimated using scores from expert raters to serve as known scores.

Rater agreement: The degree to which raters assign equivalent scores to the same domain or performance.

Rater error and systematic bias: Random and systematic variation in scores that occurs 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 that are different than those warranted by performance.

Rater-mediated assessment: An assessment that requires human scoring according to a set of criteria using a rating scale with one or more domains.

Rater monotonicity: The probability that a student will receive a higher rating increases as their location on the latent variable increases.

Rating quality: The degree to which the ratings assigned to a response are warranted by the quality of the performance.

Rater response function (RRF): The RRF describes the functional relationship between the probability that a rater assigns a score in Category *k* or higher and the overall severity for the rater in that rating scale category. The RRF may be defined using cumulative probabilities, as in the case of Mokken's nonparametric models, or using conditional probabilities, as in the Partial-Credit formulation of the Rasch model.

Raw score: A row or column total in a student-by-item (or student-by-rater) data matrix. Raw scores for students represent the sum of their scores across a set of items (or raters); raw scores for items (or raters) represent the sum of responses to the item across a group of students. Raw scores may also be called *sum scores* or *total scores*.

Rater unidimensionality*:* Ratings reflect evidence of a single latent variable. Rater unidimensionality implies that ratings are not unduly influenced by construct-irrelevant variables, such as student characteristics (e.g., gender or handwriting), rater characteristics (e.g., rating or teaching experience), or characteristics of the assessment system (e.g., prompts or assessment consequences).

Restscore: The restscore is the raw score (X_+) minus the score on an item (or from a rater) of interest. Restscores are often used in place of θ estimates in order to check nonparametric model requirements.

Scalability: The degree to which a set of items (or raters) matches the expectations of a deterministic Guttman scale (Guttman, 1950).

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

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

Unidimensionality: Item responses reflect evidence of a single latent variable.

Useful measurement properties: Measurement properties that are obtained through the use of ideal-type models, including the ability to describe persons and items on the same scale, and to obtain item- (or rater-) invariant estimates of person locations and person-invariant calibrations of item difficulties (or rater severities).

Variable map: A visual display that represents the operational definition of the latent variable and includes locations for items, persons, and other facets of interest.

Overview of Dissertation

This dissertation is organized as follows. Chapter One provided an introduction to the study including a theoretical framework, statement of the problem, the purpose of the study, and an outline of the questions guiding this research. Chapter Two addresses the first two guiding questions: 1) What are the major underlying measurement issues related to rating quality? and 2) How have these measurement issues been traditionally addressed in previous research? The chapter includes a review of literature that describes persistent concerns related to the use of rater-mediated educational assessments, and traditional methods for addressing these concerns using indicators of rater agreement, error and systematic bias, and accuracy. Chapter Three and Chapter Four present the IRT models that are used in this study and illustrate rating quality indices based on these models with an example dataset. Specifically, Chapter Three describes IRT in general, and provides a theoretical discussion and empirical demonstration of the parametric IRT models used in this study that are based on Rasch measurement theory (Rasch, 1960/1980). Chapter Four describes and illustrates Mokken scaling (Mokken, 1971), and proposes the use of Mokken's nonparametric models as methodological tools for exploring measurement quality in the context of rater-mediated educational assessments. Chapter Five is an empirical application of Rasch measurement theory and Mokken scale analysis as tools for exploring the structure of rating scales in a large-scale rater-mediated writing assessment. Finally, Chapter Six draws connections among the first five chapters and provides tentative conclusions for the guiding questions. The final chapter also includes directions for future research and a discussion of the implications of this work for research, theory, and practice.

Chapter Two: Review of Literature

In Chapter One, the first two components of the theoretical framework for this study were introduced: 1) a theory of human judgment, and 2) a theory of measurement. Specifically, a lens model for rater-mediated assessment (Figure 3) was presented based on Brunswik's (1952) lens model (Figure 2) and Social Judgment Theory (Cooksey, 1996a; Hammond and Joyce, 1975) that highlighted the influence of various cues, or mediating variables, on rater judgment. The lens model for rater-mediated assessment emphasizes the importance of considering the ecological context in which ratings are assigned, interpreted, and applied. Second, rater-invariant measurement was presented as a measurement theory in which to consider the quality of rater-assigned scores in terms of the requirements for invariant measurement. The third component of the theoretical framework for this study is evidence of rating quality. In this chapter, the concept of rating quality is introduced using a literature review that explores traditional approaches to evaluating the quality of ratings.

In this study, *rating quality* is defined as the degree to which the ratings assigned to a response are warranted by the quality of the performance. In terms of the lens model for rater-mediated assessments (Figure 3), rating quality can be conceptualized as the match between a student's location on a latent variable (θ_P) and a rater's judgment about the student's performance (θ_R). Because human raters exist within ecological contexts that mediate the relationship between θ_P and θ_R, the major underlying measurement issues related to rating quality include concerns about a rater's ability to provide a "clear reflection" of a student's performance within a particular assessment context. Focusing on the role of the rater in mediating the assessment of a student's response, Lumley (2002) summarized these concerns related to rater judgment. In his words:

> The rater, not the scale, lies at the centre of the process. It is the rater who decides:
>
> - which features of the scale to pay attention to;
> - how to arbitrate between the inevitable conflicts in the scale wordings; and
> - how to justify her impression of the text in terms of the institutional requirements represented by the scale and rater training. (p. 267)

Essentially, concerns about rating quality are related to the influence of mediating variables on rater interpretation of a performance in terms of a construct. In this chapter, the concept of rating quality is explored in terms of persistent measurement issues related to rater-mediated assessments. Specifically, a literature review is used to explore previous research on measurement issues related to rating quality. The literature review is organized around the first two research questions for the dissertation: 1) What are the major underlying measurement issues related to rating quality? and 2) How have these measurement issues been traditionally addressed in previous research? The subsequent chapters explore methods for evaluating rating quality based on parametric and nonparametric item response theory models for raters.

What are the major underlying measurement issues related to rater-mediated assessments?

The major purpose of this dissertation is to explore indices of rating quality based on parametric and nonparametric Item Response Theory (IRT) models for raters. However, before indicators of rating quality based on different models can be compared, it is necessary to consider the underlying measurement issues that have motivated the initial and continued development of quality control indicators for rater-mediated assessments. Accordingly, the first research question for this study seeks to identify the major underlying measurement issues that characterize rater-mediated assessments. In the next section, previous research is used to highlight persistent concerns related to the quality of ratings in large-scale rater-mediated assessments. Then, traditional methods for addressing these concerns are summarized using a literature review.

Previous research on rater-mediated assessments reveals a variety of concerns related to the interpretation of rater-assigned scores as meaningful descriptions of a student in terms of a construct. A common theme across previous research that focuses on the quality of ratings is concern with the subjectivity that is associated with human judgment. In general, research on the quality of rater-assigned scores suggests that consistency provides evidence of high-quality ratings; that is, evidence of consistent ratings supports their interpretation and use in high-stakes contexts. Highlighting this concern as a persistent measurement issue, research on rater-mediated assessments reveals a wide variety of methods to detect and control for inconsistent ratings. For example, Edgeworth's (1890) research that describes disagreements among judges scoring written compositions is often cited as an early example of concerns about the quality of rater-assigned scores. He described differences among human judgments in psychophysical experiments and in ratings of responses to educational tests. Observing that errors of judgment typically conform to a predictable distributional shape, Edgeworth proposed methods for evaluating the magnitude of errors in human judgment in order to inform the interpretation and use of rater-assigned scores. In his words:

> The most striking degree of discrepancy between marks which I have observed occurs in marks given at an examination in classical composition. The mark of

> one examiner is occasionally five times, and once sixteen times as great as the mark assigned by his equally and highly skilled colleagues to the same piece of Greek verse…. But it is not our part to moralise on human fallibility in general. At present our more pleasing task is to show that even in the midst of the grossest ignorance and wildest error, there may be found a drop of science if we but diligently press it out. (p. 467)

Systematic investigation of rater inconsistency continued after Edgeworth expressed these concerns related to errors in rater judgments. Nearly 50 years later, Guilford's (1936) chapter on rating scale methods in the first textbook on psychometrics echoed Edgeworth's concerns with subjectivity in human judgment. Guilford cautioned: "raters are human and they are therefore subject to all of the errors to which humankind must plead guilty" (p. 272). Later, Saal, Downey, and Lahey (1980) conducted a review of research on rater-mediated assessments and identified numerous methods for detecting and describing the influence of subjectivity and other types of judgmental errors on rater-assigned scores. Noting that different methods for monitoring rating quality reflect slightly different concerns about rater judgment, they concluded that "most of the reservations, regardless of how elegantly phrased, reflect fears that rating scale data are subjective (emphasizing, of course, the undesirable connotations of subjectivity), biased, and at worse, purposefully distorted" (p. 413).

In a recent summary of concerns about rating quality entitled *Worrying about Rating*, Hamp-Lyons (2007) identified trends in measurement research related to rating quality within the context of writing assessment. She described a persistent concern related to the "problem of the need to increase the reliability of ratings" throughout the

20th century, and observed that two major solutions are used in practice to counteract unreliability: 1) the combination of essay-based (i.e., direct) assessments with multiple-choice measures of the same construct, and 2) improved rater training that focuses on raters “as people with opinions and values” (p. 1). Similarly, Wolfe and McVay (2012) reviewed methods used in research to evaluate the quality of ratings, and observed that research in this area is characterized by two major themes: 1) descriptions of the influence of rater characteristics, such as experience, training, and prior knowledge, on the quality of ratings and 2) presentations of statistical procedures for monitoring the quality of ratings. Rating quality research based on the first theme has increased through the use of cognitive models to examine rater decision-making processes (Barkaoui, 2011), such as signal detection models (DeCarlo, 2005), the impact of rater background characteristics on rating quality (Lumley & McNamara, 1995; Pula & Huot, 1993), as well as the impact of specific types of rater training procedures (Weigle, 1998), scoring criteria (Clauser, 2000) and feedback (Elder, Knoch, Barkhuizen, & von Randow, 2005; Knoch, 2011) on rating quality. However, as Hamp-Lyons (2011) pointed out, advances in these areas are not yet sufficient to fully understand rating processes. She writes:

> The familiar interaction of contexts with tasks, texts, and raters remains at the heart of writing assessments. Despite many studies of raters and rating processes in recent years, we still do not fully understand the characteristics of raters: should they be experts or novices? Does it matter whether or not they are teachers? How much difference does rater training make? (p. 4)

Despite increased research on these background- and training-related aspects of rating procedures, concerns about rating quality are addressed in practice through methods related to statistical techniques for monitoring raters during operational scoring (Johnson, Penny, & Gordon, 2009; Wolfe & McVay, 2012).

How have these measurement issues been traditionally addressed in previous research?

In general, research on rating quality over the last century has focused on examining rating behavior within and across individual raters as a method for determining the degree to which ratings can be interpreted as an accurate reflection of a performance (Elliot, 2005; Saal, Downey, & Lahey, 1980). An examination of previous research reveals a wide range of indicators of rating quality that classify problematic rating patterns as specific types of rater errors. Saal, Downey, and Lahey (1980) provide a useful overview of the history of the classification of rater errors beginning with Thorndike's (1920) foundational work on the concept of *halo error*, or the tendency for raters to adopt a holistic view of a performance when an analytic view is warranted. Following Thorndike, Kingsbury (1922, 1933) identified three major types of rater errors: 1) leniency, 2) range restriction, and 3) halo error. With some additions, these three types of rater errors remain prevalent in research on rater errors since Kingsbury's initial classification (Murphy & Cleveland, 1991; Saal, Downey, & Lahey, 1980). For example, Guilford's (1936) chapter on rater-mediated assessments describes a set of specific patterns in ratings that are classified as different types of errors in rating related to three major categories of rating quality indices: 1) leniency and severity errors, 2) halo errors, and 3) indices of interrater reliability and agreement. Included within these categories are methods for evaluating rating quality based on measures of central tendency, correlations, factor analysis, analysis of variance, and reliability coefficients. Subsequent work in this area includes indices of rater reliability and errors, along with indices of rater accuracy—

defined as a match between ratings from operational raters and “expert” raters (Sulsky & Balzer, 1988; Murphy & Cleveland, 1991).

Current operational methods for monitoring ratings typically include indices of rating quality that are classified within three major categories: 1) rater agreement, 2) rater errors and systematic biases, and 3) rater accuracy (Murphy & Cleveland, 1991; Johnson, Penny, & Gordon, 2009). These three categories of rating quality are defined in Table 1. First, indices of rater agreement describe the degree to which raters assign matching scores to the same performance. Next, rater errors and systematic biases are used to describe specific patterns or trends in rating behavior that are believed to contribute to the assignment of scores different from those warranted by a student’s performance. Third, rater accuracy is defined in practice as a match between operational ratings and those established as “true” or “known” ratings by individuals or committees of expert raters (Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991). Accuracy is determined by comparing an observed rating to an expert-assigned rating; smaller differences between these two scores are associated with higher accuracy. When indices of rating quality based on these three categories are applied, high levels of agreement, low levels of error and systematic bias, and high levels of accuracy are assumed to reflect high-quality ratings.

In order to address the second research question for this study, the next section provides an overview of rating quality indices related to the three general categories of rater agreement, rater errors and systematic biases, and rater accuracy that are used in practice to monitor rating quality. Although indices of rater agreement, error and systematic biases, and accuracy have been examined from the perspective of several different measurement theories, such as Classical Test Theory (CTT) and IRT, the next section focuses methods for examining rating quality that are implemented in practical settings based on a "traditional" approach. Table 2 through Table 4 outline the indicators of agreement, error, and accuracy that are described in the sections below.

1. What are the major indices of rater agreement?

The first major category of traditional rating quality indices is rater agreement. As defined in Table 1, indices of rater agreement describe the degree to which raters assign matching scores to the same performance. Numerous coefficients have been proposed to evaluate rater agreement based on assumptions and conditions that underlie specific measurement situations, such as the type of rating scale, the number of rating scale categories, and the number of raters. Rather than providing a comprehensive summary of all agreement statistics that are applicable to rater-mediated assessments, this section summarizes indices of rater agreement that are routinely applied in practice as evaluative measures of rating quality. Specifically, indicators of rating quality are summarized that are applicable to rating procedures for which rating scales are composed of two or more ordered categories (i.e., polytomous ratings) and multiple raters score student responses.

In order to organize the presentation of these agreement indices, this study draws upon the theoretical classification of rater agreement coefficients presented by Stemler and Tsai (2008) into two major categories: A) indicators of *categorical agreement*, and B) indicators of *ordinal agreement*. Indices within these two categories reflect slightly different conceptualizations of rater agreement; these distinctions are highlighted in Table 2 and elaborated below. First, three categorical agreement indices (A) are described: A1) absolute agreement, A2) adjacent agreement, and A3) chance-corrected agreement. Then, four ordinal agreement indices (B) are described: B1) correlation coefficients, B2) coefficient alpha, B3) intraclass correlation coefficients, and B4) coefficients from Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). A set of statistics and displays correspond to each of these indicators of rating quality that are used in practice to detect agreement among raters; these statistics and displays are summarized in the last column of Table 2.

A. Categorical Agreement Indices

Categorical agreement indices are based on the premise that consensus among a group of raters regarding the classification of performances provides support for the existence of an underlying construct that is being measured by a group of raters. Essentially, these indices describe the degree to which pairs and groups of raters categorize performances in the same way. In operational settings, categorical agreement coefficients are often applied during rater training procedures because of their diagnostic value in identifying individual raters who may be unclear about the appropriate application of a rating scale or rubric criteria. However, practical limitations associated with the extension of these indices beyond pairs of raters and violations of assumptions of statistical independence between pairs of raters challenge the merit of these coefficients as rating quality indices beyond rater training situations (Stemler & Tsai, 2008). Categorical agreement indices that are frequently used in practice include indices of A1) absolute agreement, A2) adjacent agreement, and A3) chance-corrected agreement.

A1. Absolute agreement. Absolute agreement describes the proportion of matching ratings assigned by groups of two or more raters to the same response. For pairs of raters, the percent of matching ratings can be identified by cross-tabulating ratings assigned by the two raters and examining the proportion of shared ratings along the diagonal (Johnson, Penny, & Gordon, 2009). Measures of absolute agreement are then used to identify raters who appear to be rating in a manner inconsistent with other raters, which is viewed as a potential threat to rating quality. Describing absolute agreement statistics, Hayes and Hatch (1999) note that these indices of rating quality may be artificially inflated if most of the performances have ratings in the same category. In other words, if most students earn a rating in Category *k*, then it is likely that raters will classify these students in a similar way—thus inflating measures of agreement. Further, the practical value of absolute agreement statistics may be limited by the fact that chance agreement is not considered in the computation of percent agreement statistics.

A2. Adjacent agreement. Because it is difficult to train raters to obtain absolute agreement, indices of *adjacent agreement* can also be used to describe consensus among raters. When rating scales with more than two categories are used, adjacent agreement statistics describe agreement between two raters in terms of the proportion of ratings in adjacent categories. For example, if a student is assigned a rating in Category *k* by Rater *i*, a rating in Category $k - 1$ by Rater *j*, and a rating in Category $k - 2$ by Rater *m*, on a five-point rubric, Rater *i* and Rater *j* would be said to be in adjacent agreement and Rater *m* would not be in adjacent agreement to either of the other two raters.

Because indices of absolute and adjacent agreement are somewhat intuitive to compute and explain, these categorical agreement indices are the most prevalent agreement index used in practice to describe the match between ratings assigned by pairs of raters (Hayes & Hatch, 1999; Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991; Stemler & Tsai, 2008). Absolute and adjacent agreement statistics are often applied to monitor rating quality during operational scoring for large-scale assessments that require at least two independent ratings of a performance (Hieronymous, Hoover, Cantor, & Oberley, 1987; Wiley & Haertel, 1996; Kobrin & Kimmel, 2006). When adjacent and absolute agreement statistics are applied in practice to monitor ratings, violations of agreement are frequently resolved through the use of score resolution methods (Penny & Johnson, 2011). Usually, a discrepancy between two ratings of the same performance is large enough to be considered disagreement when the ratings are two or three score points apart (East, 2009; Hogan & Mishler, 1980; Wolcott, 1998). Several different methods for score adjudication may be employed in these instances of disagreement. These resolution methods frequently include the use of a rating from a more-experienced rater (Johnson, Penny, & Gordon, 2000; Johnson, Penny, Fisher, & Kuhs, 2003; Johnson, Penny, & Gordon, 2001; Penny & Johnson, 2011). However, different methods result in different values of resolved scores, and the decisions made based on these resolved scores have been shown to vary depending on the resolution method used (Penny & Johnson, 2011).

A3. Kappa statistics. Another consideration related to rater agreement is the fact that two raters might assign the same rating to a response simply by chance. In order to control for the influence of chance on the observed agreement between raters, Cohen (1960, 1968) proposed the use of *kappa* (κ). Kappa is an extension of Scott's (1955) chance-corrected agreement statistic for pairs of raters. Based on the assumption that pairs of ratings are statistically independent, values of the kappa statistic describe the proportion of observed agreement that is corrected for the expected level of agreement, given the marginal distributions of two raters (Banerjee, Capozzoli, McSweeny, & Shina, 1999). For pairs of raters, Cohen's (1960) kappa statistic is:

$$\kappa = \frac{P_A - P_C}{1 - P_C} \quad , \tag{1}$$

where

P_A = proportion of observed agreement, and

P_C = proportion of expected agreement based on chance.

P_C is calculated using a contingency table approach that determines the expected value within a cell, given the row and column totals. When the ratings assigned by the pair of raters can be shown to be statistically independent, a value of κ = 0 suggests that only chance-level agreement contributed to the consistency within a pair of raters (Agresti, 1992). Despite the fact that guidelines have been proposed for the interpretation of kappa (Landis & Koch, 1977), interpretation of this coefficient is somewhat complex. For example, Stemler and Tsai (2008), Agresti (1992), and others point out difficulties in comparing kappa across studies that have different base rates, and that the statistic may be better used as an overall comparison with chance agreement. Furthermore, kappa has also been criticized because the coefficient assigns equal weight to all disagreements, regardless of the magnitude of the difference (Fleiss & Cohen, 1973). Since the original presentation of kappa, the coefficient has been extended and generalized for use in a variety of rating scenarios. For example, the weighted kappa statistic (Cohen, 1968; Fleiss & Cohen, 1973) uses pre-specified weights that reflect the seriousness of disagreements for a particular rating situation. The weighted version of kappa adds a specified disagreement weight to the calculation of the P_A and P_C values in Equation 1. This is especially relevant in situations where score differences lead to different consequences, such as pass/fail decisions. Further, Fleiss (1971) generalized Cohen's (1960, 1968) kappa statistic for use with three or more raters. Banerjee, et al. (1999) provide a review of interrater agreement measures based on kappa, including weighted kappa coefficients, intraclass kappa coefficients, and kappa coefficients that include covariates, among others. They concluded that the choice of a kappa coefficient should match the assumptions regarding the underlying marginal distribution of the ratings.

B. Ordinal Agreement

Ordinal agreement indices are based on correlation coefficients. In the context of rater-mediated assessments, ordinal agreement indices describe the reliability, or consistency, in person ordering across a group of raters. Based on the classical true score model of reliability, ordinal indices of rater agreement attempt to quantify the variance within assessment situations that can be attributed to differences between raters. Specifically, reliability is defined within the framework of CTT as an estimate of the proportion of observed score variance attributable to true score variance (Crocker & Algina, 1986). In his seminal article on test reliability, Cronbach (1947) defined 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). This section continues the list of agreement indices that are used in practice to monitor rating quality. Four types of ordinal agreement indices are presented: B1) correlation coefficients, B2) coefficient alpha, B3) intraclass correlation coefficients, and B4) coefficients from Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

B1. Correlation coefficients. Correlation coefficients describe the association between ratings assigned by pairs of raters. Values of correlation coefficients range from −1 to +1, and values near |1| suggest that the ratings assigned by one rater can be used to predict the rating assigned by the second rater. Correlations near zero indicate that the ratings assigned by one rater cannot be used to predict that of the second. In contrast to categorical agreement indices, high values of correlation coefficients can be obtained when there are differences in the scores assigned by different raters, as long as the differences are systematic (e.g., a persistent difference of two score points between two raters). In practice, the Pearson product-moment correlation coefficient and the Spearman rank-order coefficient can be used to describe the association between pairs of raters (Johnson, Penny, & Gordon, 2009). As given in Stemler and Tsai (2008), the Pearson correlation for two Raters *X* and *Y* is:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X^2\right)}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y^2\right)}{N}\right)}} \quad , \qquad (2)$$

where *N* is the number of ratings.

The Pearson correlation can be applied to pairs of raters when the data from each rater are approximately normally distributed. In cases where normality is not observed, the Spearman correlation (Spearman's rho) can be calculated for rank-ordered data from two raters (Lehman, 1986). Statistical software packages can be used to compute the Spearman correlation that provide corrections for tied ranks. Because the equations used for this correction are numerous, the equation for Spearman's rho is not given here. A major limitation to the use of the Spearman correlation is the fact that this coefficient requires a fully crossed rating design (both raters score all of the performances).

B2. Coefficient alpha. A measure of agreement that can be applied to circumstances with more than two raters is Cronbach's (1951) coefficient alpha. Although it is usually applied to selected-response items, alpha can be extended to the context of rater-mediated assessments by treating raters as a sort of "item" with polytomous scores (Abedi, 1996). Coefficient alpha can be calculated as:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sigma_i^2}{\sigma_w^2}\right) \quad , \tag{3}$$

where

N = number of raters

$\sigma^2{}_i$ = variance of the ratings assigned by rater *i*, and

$\sigma^2{}_w$ = total variance of the ratings assigned across raters (variance of the total scores).

Values of coefficient alpha range from 0 to 1, and they describe the internal consistency of a set of ratings across a group of raters. Values close to 1 suggest that the majority of observed variance is due to differences in true scores. As stated by Stemler and Tsai (2008), coefficient alpha is "useful for understanding the extent to which the ratings from a group of judges hold together to measure a common dimension" (p. 39). In other words, evidence of internal consistency for a group of raters supports the claim that the raters are interpreting student performances in a similar way.

B3. Intraclass correlations. The intraclass correlation coefficient (ICC) is another measure of association that can be applied to the context of rater-mediated assessment. A variety of ICCs can be specified to match the context of a rating situation. The general form of the ICC for rating data is a ratio of within-rater variance to between-rater variance that can be stated as:

$$ICC = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \quad , \qquad (4)$$

where

$\sigma^2{}_b$ = variance of ratings between raters, and

$\sigma^2{}_w$ = pooled variance within raters.

Krippendorff (1970) and Fleiss and Cohen (1973) demonstrated that the ICC is equivalent to the weighted kappa statistic (Cohen, 1968), based on an interpretation of the mean difference between raters as a component of variability. The ICC is considered a conservative estimate of interrater reliability because this coefficient controls for unreliability related to overall (mean) differences among raters and low correlations among raters. Shrout and Fleiss (1979) presented six forms of the ICC that can be applied to the context of rater-mediated assessment, and they proposed a set of guidelines for selecting an appropriate form of the ICC as an indicator of rater reliability. They emphasized the fact that different forms of the ICC provide different information about a set of ratings, and that the choice of an ICC should be guided by the appropriate specification of a statistical model, the specification of relevant sources of error, and the purpose for conducting reliability analyses. Limitations to the use of ICCs include attenuation as a result of group homogeneity and sensitivity to non-normal distributions of rating data (Stemler & Tsai, 2008).

B4. Generalizability theory. Finally, indices of ordinal rater agreement can be obtained through the use of Generalizability Theory (G Theory; Cronbach et al., 1972;

Brennan, 2001; Shavelson & Webb, 1991). G theory procedures combine methods for estimating variance components from Analysis of Variance (ANOVA) with reliability estimation techniques from CTT in order to evaluate the consistency, or reliability, of a set of observations within a particular context (Brennan, 2001). Within the framework of G theory, information about the reproducibility of a set of scores is viewed as essential for their interpretation and use. Methods based on G theory expand the classical approach to reliability analyses by allowing the researcher to specify a set of conditions that are of interest for a particular measurement procedure, and to examine the impact of each of these conditions on the variance in observations. Specifically, G theory is used to partition an observation into an effect for the object of measurement, an effect for each additional facet, or source of variance, and an effect for each of their combinations. Using ANOVA, variance components are estimated for each facet and interaction. The variance components are then used to identify major sources of measurement error, estimate the total magnitude of the measurement error, and form a reliability coefficient; this set of techniques is called a G study, and it is the first step in the two-stage process of G theory analyses. Following a G study, the researcher can explore the impact of adjustments to sample sizes for various conditions through analyses termed *D studies*, or decision studies. Essentially, D studies allow the researcher to explore the impact of various sources of error on measurement reliability under different conditions, or *universes of generalization*. These specifications reflect designs to which results from the current study are to be generalized. Results from D studies include reliability coefficients and estimates of the standard error of measurement that can be used to inform assessment development and implementation procedures.

Several authors, including Brennan (1996, 2000), Cronbach, Linn, Brennan, and Haertel (1997), Huang (2008), Johnson, Penny, and Gordon (2009), and Lane and Stone (2006), have demonstrated the application of G theory as a measurement framework in which to consider the relationship between the quality of ratings and a variety of aspects of rater-mediated measurement procedures, including raters, tasks, and schools. G theory analyses are useful in the context of performance assessment because they allow researchers to specify analysis designs that represent a particular measurement situation. For example, facets such as raters or rubrics may be considered fixed or random for different measurement purposes (Brennan, 1996). The variance components that are estimated in a G study can be used to identify aspects of a rater-mediated assessment system that may contribute to variation among raters. Within the G theory framework, rater reliability is described as "rater accuracy." However, accuracy from the perspective of G theory is a distinct concept from the rater accuracy indices described below. In addition to variance components, G theory analyses of rating quality also focus on the standard error of measurement as an indicator of rater precision. Cronbach, Linn, Brennan, and Haertel (1997) highlighted the importance of the clear specification of facets in G theory designs for the interpetation of standard errors. Similarly, Brennan (1995) emphasized the fact that estimates of precision based on G theory can be adapted to describe a variety of measurement purposes. In his words:

> Strictly speaking, there is no such thing as *the* standard error of measurement. There are numerous possible standard errors of measurement corresponding to different universes and designs. Indeed, standard errors of measurement often can be made arbitrarily large or small by broadening or narrowing the universe.

> Therefore, statements about standard errors of measurement should not be judged in the abstract but should be interpreted relative to a clear specification of the universe. (p. 273, italics in the original)

This flexibility necessitates the consideration of the G theory design when interpreting standard errors as indices of measurement quality.

Summary of Agreement Indices

The above discussion highlighted popular methods for examining rater agreement within two major categories: A) categorical agreement, and B) ordinal agreement. When interpreting indices of categorical and ordinal rater agreement, it is important to consider the different types of information provided by coefficients within these two categories. Although they are both classified as types of agreement indices, indices of categorical agreement emphasize rater exchangeability, while ordinal agreement indices describe the relative consistency of a group of raters (LeBrenton & Senter, 2008). Reviews and summaries of rater agreement measures are widespread that include the categorical and ordinal agreement coefficients described above, as well as additional indices of rater agreement (e.g., Banerjee, 2006; Burry-Stock, Shaw, Laurie, & Chissom, 1996; Hayes & Krippendorff, 2007; Shoukri, 2010; Uebersax, 1992, 2002; Zegers, 1991). In a review of performance assessment research reporting measures of rater agreement, Jonsson and Svingby (2007) noted that, despite the plethora of agreement statistics that can be applied to the context of rater-mediated assessment beyond those presented here, most evaluations of rating quality using agreement indices focus on indicators of categorical and ordinal agreement. Specifically, these authors found that most studies in which categorical agreement was reported included estimates of absolute agreement and adjacent agreement statistics, with most studies reporting between 55% and 75% exact agreement and over 90% adjacent agreement. Studies that reported ordinal agreement generally did not specify which estimate of rater consistency was calculated; however, when the coefficient was specified, Pearson, Spearman, and Kendall's *W* coefficients (Kendall, 1938) were most frequent.

Along with concerns related to the broad range of methods for calculating interrater agreement, conceptual issues challenge the validity of agreement statistics as indices of rating quality. In their discussion of error and accuracy measures for performance appraisal ratings, Murphy and Cleveland (1991) described an inherent conflict in the interpretation of interrater reliability measures. They discussed 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, Lumley (2002) described difficulty in interpreting reliability estimates within the context of large-scale writing assessments of English as a Second Language because of the many factors that influence ratings. In his words: "Levels of reliability are relatively easy to calculate. What is less clear is what the basis of the ratings actually is: how can we account for this host of other factors?" (p. 249). These authors suggest that a simple measure of consistency among ratings may reflect agreement 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.

2. What are the major indices of rater errors and systematic biases?

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 specific trends or patterns in ratings. 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, rater errors and systematic biases can be defined as 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 previous research reveals a wide range of definitions for rating errors and biases, with an equally wide range of methods for classifying these rating patterns. As described above, the systematic classification of rater errors can be traced to research during the early 20th century on halo errors (Thorndike, 1920), errors related to restricted use of the rating scale, and errors related to rater severity and leniency (Kingsbury, 1922, 1933). On the other hand, systematic bias in ratings is conceptualized in a similar fashion to bias in multiple-choice assessments. As defined by Cole & Moss (1989): "Bias is differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (p. 205). Although rater errors and systematic biases reflect distinct concerns about the quality of ratings, both are related to patterns in the use of scoring rubrics that result in predictably higher or lower scores than are warranted by a response. Research that examines potential causes for the presence of rater errors and systematic biases generally links these phenomena to individual raters' desire or pressure to match the rest of a group, individual rater differences in interpretation of a rating scale, or systematic biases that result in high or low ratings on performances with certain construct-irrelevant characteristics, such as length, handwriting, or gendered language (Lane & Stone, 2006; Murphy and Cleveland, 1991; Johnson, Penny, & Gordon, 2009).

Table 3 summarizes indices of rater error and systematic bias that have been traditionally used to monitor the quality of ratings in large-scale educational assessments. The next section summarizes methods for identifying rating patterns that can be categorized as A) distributional errors: A1) errors of leniency and severity, A2) range restriction, and A3) central tendency; B) correlational errors: B1) halo error; or C) systematic biases: C1) interactions. Similar to the rater agreement indices presented earlier, a variety of statistics and displays correspond to these indices of rater errors and systematic biases. Table 3 lists these methods for detecting rater errors and systematic biases.

A. Distributional Errors

The first major category of rater errors is distributional errors. Rating patterns that are classified as distributional errors are based on assumptions about the underlying distribution of true scores within a population of interest. Essentially, distributional errors describe a mismatch between an observed distribution of ratings and the assumed underlying true score distribution. Murphy and Cleveland (1991) called attention to the limitations of the assumptions underlying the definition of these phenomena as errors. Expressing concern with the operational use of these rating patterns as indicators of rating quality, they observed:

> There are two reasons to be concerned about the use of distributional error measures to infer that ratings are accurate or inaccurate. First … the true distribution of performance is almost never known. Indeed, if there were means available to determine the true distribution of performance, it is hard to see why ratings would be needed at all. It is doubtful that anyone would favor the use of subjective criteria *if valid objective criteria were available.* They typically are not, meaning that the assumptions that underlie distributional error measures are inherently untestable**.** We believe that they are also implausible. Second, ratings whose distributions *did* correspond to the (unknown) true distribution of performance are not necessarily more accurate than those whose distributions are obviously wrong. (p. 219, italics in original)

Despite these concerns, distributional errors are used in practice to monitor rating quality. Three types of distributional errors that are frequently used include: A1) errors of leniency and severity, A2) range restriction, and A3) central tendency.

A1. Leniency and severity. The first type of distributional error is rater leniency and severity. Although there are a variety of definitions for rater leniency and severity errors, generally accepted definitions are as follows: Raters are assumed to be *lenient* when their average ratings are systematically higher than those assigned by the rest of a rater group; Raters are assumed to be *severe* when their average rating is systematically lower than those assigned by the rest of a rater group. Indicators of rater severity and leniency include a comparison of average ratings for individual raters with average ratings across a group of raters, examination of rater main effects in a rater-by-student-by-domain ANOVA, and examination of the skewness of a rating distribution (Murphy & Cleveland, 1991; Saal, Downey, & Lahey, 1980).

A2. Range restriction. Next, the rater error of *range restriction* refers to a rater's tendency to assign ratings that cluster around a particular rating scale category; this category may be located anywhere on the rating scale. Essentially, the definition of this rater error reflects a view that the true scores in a population are distributed across the score range, such that a uniform or tightly clustered rating distribution would be incorrect. Indices of range restriction that are used in practice include small standard deviations for individual raters across students within domains, kurtosis of a rating distribution, and the lack of a significant student main effect in a rater-by-student-by-domain ANOVA (Murphy & Cleveland, 1991; Saal, Downey, & Lahey, 1980). The lack of a student main effect within domains may provide evidence that raters are not detecting meaningful differences among student performances.

A3. Central tendency. The third type of distributional error is central tendency. *Central tendency error* is a type of range restriction that describes a rater's tendency to assign scores near the midpoint of a rating scale. Citing DeCotiis (1977), Saal, Downey and Lahey (180) described central tendency as "a rater's unwillingness to go out on the proverbial limb in either the favorable or unfavorable direction" (p. 417). Although it is a type of range restriction, the fact that central tendency describes a clustering of ratings near the midpoint of a rating scale establishes this rating pattern as a distinct rater error. Common approaches for identifying central tendency are similar to those that are used to recognize range restriction. However, examination of the proximity of average ratings within a domain to the midpoint on a rating scale is also used in practice to identify central tendency (Saal, Downey, & Lahey, 1980). When the majority of ratings are close to the center of the scale, a rater may be demonstrating central tendency.

B. Correlational errors

Next, correlational errors focus on raters' ability to distinguish among distinct aspects of a performance, such as the meaning and mechanics domains on a rubric for a writing assessment. Saal, Downey, and Lahey (1980) described correlational errors as the result of a rater's tendency to score an entire performance based on only one domain when there are actually multiple distinct domains to be scored, or as the result of conceptual similarities across distinct domains that result in a rater's inability or unwillingness to discriminate among aspects of a performance. Research on these inflated domain intercorrelations suggests that conceptual similarity across domains may be augmented by the influence of a variety of factors during a rating process, including monitoring procedures that specify "hit rates," or required frequencies for ratings within particular rating scale categories, systematic biases related to characteristics of a performance, and the tendency for raters to discount inconsistent information in a response (Cooper, 1981; Saal, Downey, & Lahey, 1980). Correlational error is among the earliest classifications of specific types of rater patterns as errors (Bingham, 1939; Thorndike, 1920), and it is usually described as *halo error.*

B1. Halo error. The term *halo error* was introduced by Thorndike (1920) to describe situations in which the global evaluations of a performance affect the evaluation of specific aspects of a performance; this definition of halo error is sometimes referred to as a *strong interpretation* of halo error, and its presence is difficult to support without the use of controlled experiments (Nisbett & Wilson, 1977; Saal, Downey, & Lahey, 1980).

However, halo error is also used to describe situations in which ambiguous information about a construct may limit a rater's ability to distinguish among domains; this interpretation is referred to as a *weak interpretation* of halo error (Saal, Downey, & Lahey, 1980). A prevalent theme in research on halo error is related to the fact that observed correlations among domain ratings may reflect true correlations among these domains within performances. This concept of *true halo* as a distinct phenomenon from *illusory halo*, or halo that actually reflects erroneous rating patterns, was introduced by Bingham (1939). Similar to issues with distributional errors, it is difficult to distinguish between true and illusory halo in practice, and indices of halo error are based on assumptions related to the "true" distinctiveness of a set of domains. Murphy and Cleveland (1991) and Saal, Downey, and Lahey (1980) identified several methods for identifying halo error in practice. Specifically, evidence for a halo effect may be obtained through examination of intercorrelations among domain ratings, with high correlations suggesting a lack of discrimination among domains. In addition, small standard deviations across domains and the presence of a significant rater-by-student interaction in a rater-by-student-by-domain ANOVA may suggest the presence of a halo effect. Factor analysis of a domain correlation matrix may also be used to identify halo; the percentage of variance accounted for by the first principal component may point towards dependence across domains.

C. Systematic Biases

The last category of rater errors and systematic biases includes rating patterns that suggest interactions between ratings and construct-irrelevant characteristics of assessments and students. Bias was defined earlier in this chapter using the definition from Cole & Moss (1989): “Bias is differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers” (p. 205). The presence of systematic bias in ratings has primarily been explored through the examination of rating quality within and across student populations and performance tasks, along with the use of interaction analyses to identify differences in rater severity related to certain construct-irrelevant characteristics.

C1. Interactions. In practice, indicators of systematic bias in ratings include interaction effects in ANOVA models. Specifically, the magnitude of interaction effects provides information about the degree to which ratings are invariant over internal and external components of an assessment system. When systematic biases are present, the meaning of ratings is not comparable across aspects of the assessment system; these interactions may be related to a variety of facets in an assessment system. First, interactions may be related to internal components of an assessment. *Internal components* are variables related to the assessment procedure itself, such as prompts. For example, a rater who demonstrates systematic bias related to an internal component might be systematically more severe on prompts that call for a persuasive response. On the other hand, interactions may be related to external components of an assessment system. *External components* are variables that are not related to the assessment procedure, such as student gender or handwriting. For example, a rater who demonstrates systematic bias related to external components might be systematically more severe when scoring responses composed by male students.

Summary of Error and Systematic Bias Indices

Research on rater error and systematic bias highlights a need for consistent definitions of and identification methods to identify categories of error and bias. Noting different implications for rater errors when they are inconsistently defined, Saal, Downey, and Lahey (1980) described a "lack of congruency between conceptualization and quantification" of these phenomena (p. 423). Further, Murphy and Cleveland (1991) described rater errors as indirect indicators of rating quality, and noted inconsistency between the definitions of rating errors and methods for identifying them, along with inconsistency in rater error indices over time. An observed lack of correlation among rating errors challenges their meaningful application in practice. Rather than providing useful methods for monitoring rating quality, they concluded that inconsistency in the information provided by distributional and correlational rater errors complicates the evaluation of ratings. In their words: "Our overall conclusion is that rater error measures should be abandoned. They are based on arbitrary and often implausible assumptions, and there are too many nonequivalent definitions of each one" (p. 226).

3. What are the major indices of rater accuracy?

Rater accuracy is the third category of traditional rating quality indices. Theoretically, an observed rating is accurate when it matches a student's true score for a given performance. However, true scores are generally not known, and there is no agreed-upon method for identifying them. Because these true scores are not attainable, research on rating quality uses indices of rater agreement and rater errors and systematic biases as *indirect* indices of rater accuracy, and indices of a match between observed and true scores as *direct* indices of rater accuracy (Murphy & Cleveland, 1991; Johnson, Penny, & Gordon, 2009). The indirect approach for examining rater accuracy associates high levels of agreement and a lack of errors with high-quality ratings. In contrast, 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. Another approach to defining true ratings is the use of the arithmetic average of operational ratings for a particular student as a criterion against which to compare individual raters (Wolfe & McVay, 2012). Sulsky and Balzer (1988) reviewed research between the 1970s and 1980s that used indices of rater accuracy as a method for monitoring rating quality, and recognized a common conceptualization and operational definition of accuracy as " a comparison of the rater's ratings with the true scores of ratee performance" (p. 498). Specifically, Sulsky and Balzer identified rater accuracy as "a term used to describe both the strength and kind of relation between one set of measures and a corresponding set of measures (e.g., true scores) considered to be an accepted standard for comparison (Guion, 1965)" (Sulsky & Balzer, 1988, pp. 497-498). Similarly, Woehr and Huffcutt (1994) claimed that the evaluation of rater accuracy requires a "comparison of an individual rater's rating across performance dimensions and/or ratees with corresponding evaluations provided by expert raters (i.e. 'true score'). With these measures then, the closer the raters' ratings are to the 'true scores', the more accurate they are believed to be" (Woehr & Huffcutt, 1994, p. 92). This conceptualization of accuracy as a function of the difference between operational and expert ratings is evident throughout research on rating quality. As pointed out by Berkowitz-Jones (2007), methods used in practice to estimate rater accuracy typically include three indices that can be classified into two major categories: A) categorical accuracy, and B) ordinal accuracy. First, indicators of categorical accuracy include A1) distance accuracy and A2) accuracy components. Second, ordinal accuracy is indicated by B1) correlational indices of differential accuracy. These accuracy indices are summarized in Table 4.

A. Categorical Accuracy

The first major category of rater accuracy indices is categorical accuracy. Similar to indices of categorical agreement, accuracy indices within this category focus on the degree to which operational and expert raters classify responses in the same way. In this study, two types of categorical accuracy are described. Indices of distance accuracy (A1) describe the distance between operational and expert ratings on the score scale, and indices based on accuracy components (A2) describe the correspondence between operational and expert ratings in terms of a set of components.

A1. Distance Accuracy. Consistent with Sulsky and Balzer's (1988) definition, the most common procedure for estimating rater accuracy in performance assessment literature involves an estimation of the distance between operational ratings and "true" or known ratings that are assigned by an expert rater. An early method developed for the estimation of distance accuracy is Cronbach's (1955) D^2 index, which describes the squared distance between operational and known ratings averaged across students and domains in an analytic rubric (Sulsky & Balzer, 1988, p. 498). D^2 can be expressed mathematically as:

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

where

x = observed ratings,

t = true scores,

n = students, and

k = domains.

The D^2 statistic is mathematically equivalent to the percent accuracy agreement statistic presented by Johnson, Penny, and Gordon (2009). The percent accuracy agreement statistic is an index of the absolute agreement between an operational rater and an expert rater that can be used to monitor rating quality during operational scoring. This statistic is equivalent to an absolute agreement measure between an operational and expert rater. Contrasting accuracy agreement with other measures of rating quality, Johnson, Penny, and Gordon claimed that this index "introduces the validity of rater scores that is absent from measures of interrater agreement" (p. 235). They also point out the useful application of this index with computer-based rating systems, in which percent accuracy agreement among a group of raters can be continuously monitored throughout the scoring process when an expert rater is available. Although this method can be applied in practice when expert ratings are available, Sulsky and Balzer (1988) note drawbacks related to this method for monitoring rating quality, including difficulty in interpreting values of D^2. Further, as noted by Cronbach (1955), this index may collapse potentially meaningful information about rater accuracy.

A modified version of the D^2 index is Distance Accuracy (*DA*). Sulsky and Balzer (1988) define *DA* as "the average absolute deviation of subject ratings from the true scores" (Sulsky & Balzer, 1988, p. 499). The *DA* statistic is expressed mathematically as:

$$DA_k = \frac{\sum_{j=1}^{n} \frac{(\sum_{i=1}^{d} | t_{ij} - r_{ijk} |)}{d}}{n} \quad , \tag{6}$$

where

k = Rater *k*,

n = number of students,

i = Student *i*,

d = number of domains,

j = Domain *j*,

r = observed rating, and

t = true scores.

Unlike the D^2 statistic, the *DA* statistic incorporates the magnitude of the difference between observed and expert scores.

A2. Accuracy Components. Most large-scale rater-mediated assessments involve raters rating the performance of numerous students across a range of domains and tasks. As a result, the definition of accuracy across these aspects of a rater-mediated assessment system becomes complex (Murphy & Balzer, 1989). Within these multi-faceted assessment contexts, the value of distance accuracy statistics may be limited by the fact that these statistics provide an overall evaluative index of rater accuracy that may ignore components of the assessment system. Cronbach (1955) proposed a solution to this problem by decomposing the D^2 index into a set of accuracy components using techniques based on ANOVA. Similar to variance components in ANOVA, each component of accuracy expresses a different aspect of the distance between observed and expert ratings (Sulsky & Balzer, 1988). There are four accuracy components: 1) elevation, 2) differential elevation, 3) stereotype accuracy, and 4) differential accuracy.

High accuracy on one component does not necessarily imply that a rater will be highly accurate on another component (Cronbach, 1955). Accordingly, it is important to consider each individual component to gain a more complete assessment of rater accuracy. First, *elevation accuracy* is based on operational use of a rating scale, and the degree to which the average observed rating for a single student over all tasks or domains matches the corresponding average expert rating. Second, *differential elevation accuracy* describes the difference between the observed ordering of a group of students in terms of observed total scores and their ordering based on expert scores. A rater with high differential elevation accuracy would rank-order the performance of all students from best to worst in the same way as the expert rater. Third, *stereotype accuracy* describes the match between operational and expert raters' discrimination of student performance across domains over an entire group of students. A rater with high stereotype accuracy would match the expert rater in identifying the difficulty ordering of domains across all students. Finally, *differential accuracy* describes the match between operational rater and expert rater domain ordering for individual students. A rater with high differential accuracy would match the expert rater in terms of identifying individual strengths and weaknesses. In addition to Cronbach (1955), several sources provide mathematical formulas and elaborated definitions for these accuracy components including Sulsky and Balzer (1988) and Murphy and Balzer (1989). Empirical findings reveal that there is minimal correlation between measures of rater accuracy from these four components; these findings provide further support for Cronbach's (1955) claim that rater accuracy is a multidimensional construct (Borman, 1977).

B. Ordinal Accuracy

The second category of rater accuracy indices is ordinal accuracy. Similar to ordinal agreement indices, this conceptualization of rater accuracy focuses on the degree to which operational and expert raters order student responses in the same way. Correlational measures of accuracy (B1) are used to describe this aspect of rater accuracy in practice.

B1. Correlational measures of differential accuracy. Despite the fact that some researchers acknowledge rater accuracy as multidimensional, others believe that it is neither necessary nor practical to measure all the components of rater accuracy. For example, Borman (1977) was particularly interested in the ability of raters to rank individuals on a particular attribute within the context of rater judgments of job performance. Like Cronbach's formula, Borman's index of differential accuracy (Borman's *DA*) relies on the correlation between observed and expert ratings, but the correlational components from Borman's differential accuracy and Cronbach's formulation are not equivalent (Sulsky & Balzer, 1988). Borman calculated differential accuracy as follows (Sulsky & Balzer, 1988, p. 499):

$$\text{Borman's } DA = \frac{1}{d \sum_{j=1}^{d} (T_{rt}^{*})}, \tag{7}$$

where

d = number of domains, and

T_{rt}^{*} = correlation between observed ratings and expert ratings for a particular domain, transformed to a Z-score.

Because Borman's formulation of differential accuracy is based on correlations rather than distance, Sulsky and Balzer (1988) claimed that this measure should not be considered a direct measure of rater accuracy.

Summary of Accuracy Indices

The last section introduced rater accuracy indices within two major categories: A) categorical accuracy, and B) ordinal accuracy. In general, reviews and summaries of rater accuracy indices identify a lack of correlation between these traditionally defined rater accuracy indicators. Because each rater accuracy indicator provides distinct information, these reviews highlight the need for a contextualized interpretation of rater accuracy and the use of multiple indicators to evaluate rater accuracy in operational settings (Murphy & Balzer, 1989; Murphy & Cleveland, 1991; Sulsky & Balzer, 1988). Further, limited availability of expert raters prevents the widespread application of these rating quality indices in operational settings. Although no significant correlations have been observed between traditional indicators of indirect and direct rater accuracy, recent research has demonstrated a correlation between these two categories when the indicators are calculated using models from Rasch measurement theory (Wind & Engelhard, 2012, 2013).

Summary

In this chapter, a literature review was used to explore evidence of rating quality, which is the third component of the theoretical framework for this study (Figure 1). First, the major underlying measurement issues related to rater-mediated assessments were considered through an examination of previous research on rating quality. In general, this research identified persistent concerns related to the influence of the inconsistency and subjectivity that is often associated with human judgment on the quality of rater-assigned scores. In terms of the lens model presented in Chapter One (Figure 2), these underlying concerns are based on the influence of construct-irrelevant mediating variables, which may distort a rater's perception (θ_R) of a student's performance in terms of the construct (θ_P).

In the second part of this chapter, research that documented previous methods for addressing these concerns was reviewed. Specifically, literature was reviewed in order to provide an overview of methods that are applied in practice for monitoring rating quality based on 1) rater agreement, 2) rater error and systematic biases, and 3) rater accuracy. In terms of the lens model, rating quality indicators within these three categories represent the traditional approach to estimating the match between θ_P and $\theta_{R.}$ The wide range of rating quality indices within and across these three categories suggests that the choice of rating quality indices may influence conclusions about the quality of ratings. In general, summaries of rating quality indices emphasize the need for increased precision in the description of rating quality in order to use them as evidence to inform the interpretation and use of ratings (Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991). Drawing similar conclusions, Saal, Downey, and Lahey (1980) asserted:

> No longer can we be fuzzy in our definition of leniency, for example, and then proceed to quantify the phenomenon with three different, noninterchangeable techniques. No longer can we define halo in terms of a particular rater's behavior and then proceed to quantify the phenomenon by aggregating data collected from a group of raters. (p. 426).

In addition to the lack of precision in the definition of rating quality indices, the use of these "traditional" rating quality indices within the categories of agreement, error and systematic bias, and accuracy is further challenged by the fact that these indicators do not provide consistent information about the overall quality of a set of ratings. Reviews and meta-analyses of rating quality indices based on the categories of rater agreement, error and systematic bias, and accuracy reveal that indices across these categories do not provide consistent information about a set of raters. In their review of rating scale methods in psychological research, Saal, Downey, and Lahey (1980) demonstrated a lack of consistency among definitions of rating quality indices; they claimed that this lack of consistency increases the already subjective nature of ratings for three main reasons:

> First, there is less than unanimous agreement regarding conceptual definitions for several of the criteria of rating quality. Second, there is even less agreement regarding the operational definitions (Downey & Saal, [1978]). Third, different researchers have used different research designs or data collection procedures with inherently limited capabilities of aggregating and yielding particular statistical indices of rating quality. It is therefore easy to find two or more studies in the literature that use the same label for a particular criterion of rating quality (e.g., *halo*) even though the conceptual and operational definitions of that particular rating error are not identical and the data collection strategies are sufficiently different to preclude calculation of similar statistical indices. (p. 414)

As a result, it is not possible to generalize information about rating quality across these three categories; for example, evidence of rater agreement does not imply a lack of halo error or rater accuracy. The lack of alignment across these indices of rater agreement, error, and accuracy is further highlighted by the fact that rating quality within one category is likely to influence interpretation of results from rating quality analyses within another category. For example, because ordinal indices of rater agreement based on correlation coefficients are only sensitive to the consistency of rater rank-ordering of performances, the interpretability of interrater reliability estimates based on correlation coefficients is influenced by rater leniency and severity error (Cronbach, Linn, Brennan, & Haertel 1997; Lane & Stone, 2006; Zhu & Johnson, 2013).

Rating Quality from the Perspective of Invariant Measurement

In contrast to the disparate set of methods for examining rating quality based on indices of rater agreement, error and systematic bias, and accuracy, it is also possible to examine the quality of ratings from the perspective of invariant measurement. As discussed in Chapter One, rater-invariant measurement requires that the measurement of student achievement should not depend on the particular raters who happen to be used for the measuring. Likewise, rater-invariant measurement requires that estimates of rater severity not depend on the particular students that they score. Models based on invariant measurement can be used to provide a coherent set of rating quality indices that describe the degree to which a set of ratings meets the requirements for rater-invariant measurement. Specifically, models exist within both parametric and nonparametric IRT that can be used to examine the degree to which polytomous ratings meet the requirements of invariant measurement. Unlike traditional methods for examining rating quality that are based on a sample-dependent total score, methods for examining rater-mediated assessments based on IRT use probabilistic nonlinear models to explore the relationship between characteristics of ratings and persons (Hambleton & Jones, 1993). In the remaining chapters, data from rater-mediated assessments are examined from the perspectives of nonparametric and parametric IRT. Rasch Measurement Theory (Rasch, 1960/1980) is the parametric IRT framework, and Mokken Scale Analysis (Mokken,

1971) is the nonparametric IRT framework. Rasch- and Mokken-based models are based on a similar set of underlying requirements, and invariance can be examined within both frameworks. The next two chapters provide a theoretical overview of Rasch- and Mokken-based measurement models for rater-mediated assessments, and an example dataset is used to illustrate indices of rating quality based on the two approaches.

Chapter Three: Illustration of Modern Rating Quality Indices based on Rasch Measurement Theory

Chapter Three and Chapter Four continue to explore the third component of the theoretical framework for this study: evidence of rating quality. Specifically, these two chapters extend the traditional rating quality indices described in Chapter Two to a set of modern rating quality indices based on the two major measurement theories used in this study: Rasch measurement theory (Rasch, 1960/1980) and Mokken scaling (Mokken, 1971). Both Rasch measurement theory and Mokken scaling can be described within the framework of Item Response Theory (IRT). In terms of the lens model (Figure 2), these measurement theories provide systematic methods for evaluating the match between a rater's perception of a student in terms of a construct (θ_R) and their unobservable ("true") location on the construct (θ_P). Specifically, models and procedures based on Rasch measurement theory and Mokken scaling can be used to examine operational ratings in terms of the requirements for rater-invariant measurement that were described in Chapter One.

In this chapter, a brief definition of IRT is presented, and the major components of Rasch measurement theory are described. Chapter Four presents the major components of Mokken scaling. The presentation of each measurement approach is organized as follows: First, the original formulation of the measurement model for dichotomous data is presented theoretically and mathematically. Then, extensions of the models for use with polytomous data are presented and previous applications of these models to rater-mediated assessments are summarized. Finally, illustrative analyses with an example data set demonstrate the application of these measurement models as tools for evaluating rating quality.

The illustrative data analyses in Chapter Three and Chapter Four are secondary analyses of a dataset that was previously examined by Andrich (2010), and by Gyagenda and Engelhard (2009). The data come from the Georgia High School Writing Test, and include scores from 365 eighth-grade students whose persuasive essays were rated by 20 operational raters. Each essay was also scored by a group of expert raters called the *validity committee*. The validity committee was composed of a group of raters whose expertise was considered sufficient for the assignment of scores that reflect "true" or "accurate" measures of a student's achievement in terms of the rubric for the Georgia High School Writing Test. The scores assigned by the validity committee are used as criteria for the evaluation of the quality of scores assigned by operational raters. The ratings were assigned using a four-point rating scale (1 = *low* to 4 = *high*) and an analytic rubric with four separate domains: Conventions, Organization, Sentence Formation, and Style. All 20 operational raters and the validity committee scored the entire set of 365 essays, such that the rating design was fully connected (Eckes, 2009; Engelhard, 1997), and each essay received 21 ratings. For this study, the ratings were recoded to 0 = *low*; 3 = *high* for analyses. The Facets computer program (Linacre, 2010) is used to conduct data analyses based on Rasch models. In Chapter Four, the *mokken* package for the *R* statistical software program (*R* Development Core Team, 2013; van der Ark, 2013) is used for the analyses based on Mokken scaling.

What is Item Response Theory?

Item Response Theory (IRT) is a measurement theory that describes relationships among persons, items, and latent variables. Numerous models and methods exist within IRT that are used to explore these relationships and make inferences about persons, items, and latent variables in the social, behavioral, and health sciences (Engelhard, 2013). Using mathematical models, the major goal of IRT analyses is to predict responses based on information about persons or items. de Ayala (2009) defines IRT as follows:

> IRT is, in effect, a system of models that defines one way of establishing the correspondence between latent variables and their manifestations. It is not a theory in the traditional sense because it does not explain why a person provides a particular response to an item or how the person decides what to answer (cf. Falmagne, 1989). Instead, IRT is like the theory of statistical estimation. IRT uses latent characterizations of individuals and items as predictors of observed responses. (p. 4)

The distinguishing feature of IRT from its counterpart, Classical Test Theory (CTT), is the modeling of item and person characteristics on a single continuum that is assumed to represent a latent variable. Although most IRT models were initially developed for use with dichotomously scored selected-response (i.e., multiple-choice) items, numerous generalizations of these models are suitable for use with polytomously scored items, such as those used in many rater-mediated assessments (Engelhard, 2005). IRT serves many purposes. For example, IRT analyses can be used to score assessments or surveys, compare different assessments or surveys using a common metric, and revise or develop assessment or survey instruments (DeMars, 2010).

In order to explore relationships among respondents, items, and latent variables, IRT models compare the structure of response data to statistical models that specify expected relationships among these variables. When there is a close match between the properties of a model and the characteristics of observed data, IRT models can be used to predict responses using information about respondents and items. In order to determine whether or not an IRT model can provide useful predictions for a particular dataset, it is necessary to consider the underlying properties of the model. All IRT models are based on assumptions or requirements about the relationships among persons, items, and latent variables. Differences in these assumptions distinguish IRT models from one another, and these differences justify the use of a model for a specific purpose. It is important to recognize that there is a philosophical difference between the concepts of *model assumptions* and *model requirements* to describe the underlying principles for a particular measurement model. When a model is viewed as an ideal type (see Chapter One), the term *requirement* is more appropriate because it highlights the prioritization of the model over the data, which may or may not conform to the requirements for invariant measurement. On the other hand, the term *assumption* is appropriate when the goal of a measurement procedure is to describe the unique characteristics of a dataset using the best-fitting model. In this study, the term *requirement* will be used in reference to Rasch models (Rasch, 1960/1980), and the term *assumption* will be used in reference to models based on Mokken scaling (Mokken, 1971).

The first two columns of Table 5 list three major categories of model requirements or assumptions that are common across all IRT models. These general categories provide a framework in which families of IRT models can be compared. The first category is *dimensionality*. Dimensionality is based on the idea that response data reflect one or more person-related latent (i.e., underlying) variables. For example, a latent variable that is measured in an educational achievement test might be writing achievement. Many popular IRT models are based on unidimensionality, which requires that response data are manifestations of a single latent variable. In contrast, multidimensional IRT models describe response data as the result of two or more distinct latent variables.

The second category includes requirements or assumptions about the *independence of item responses*. Within the context of IRT, the term "item" is used broadly to represent assessment opportunities. For example, these may include selected-response items, constructed-response items, or raters who score a performance. Usually, item responses are required or assumed to be conditionally independent. Essentially, conditional independence means that a person's response to one item does not influence their response to another item, after controlling for the latent variable.

Finally, IRT models are based on requirements or assumptions about the *functional form* of the relationship between latent variables and item responses. The functional form is specified through a mathematical model that describes the probability for an observed response, given the location of a person and an item on the latent variable. This final category is particularly important for the model comparisons in this study, because differences related to the functional form assumption distinguish parametric IRT models from nonparametric IRT models. The major difference between parametric and nonparametric IRT models is related to the restrictions placed on the shape of the functional form that describes the probabilistic relationship between person locations on the latent variable and item responses. Generally, the distinction between the terms *parametric* and *nonparametric* in statistics refers to the distribution, or shape, of a variable (Lehmann, 1986; Siegel, 1956). Based on this definition, most IRT models would be considered nonparametric, because they do not place requirements on the distribution of latent variables. However, in IRT the distinction between parametric and nonparametric models refers to the functional form that underlies a particular model. Specifically, parametric IRT models specify the functional form using a specific algebraic formula, while nonparametric IRT models place only order restrictions on the functional form (Sijtsma, 1998). These differences are further explored in this and the next two chapters of the dissertation.

What is Rasch Measurement Theory?

The parametric analyses used in this study are based on Rasch measurement theory (Rasch, 1960/1980). Rasch (1960/1980) developed a probabilistic model that meets the requirements for invariant measurement; this model has been widely applied to the context of educational achievement tests. The Rasch model can be described as an IRT model because it describes the relationship between a person's location on a latent variable and the probability for an observed response. However, the development and underlying philosophy of Rasch models has motivated many of its proponents to describe it as fundamentally different from other commonly used IRT models, such as the two parameter logistic model (2PL) and three parameter logistic model (3PL; Andrich, 2004; Thissen & Orlando, 2001). As Andrich (2004) points out, major philosophical differences between models based on Rasch measurement theory and other IRT models are related to the contrasting perspectives about the role of an item response model. Proponents of the

Rasch model view the role of an item response model as an ideal structure that can be used to identify whether a set of data has useful measurement properties; as a result, the Rasch model is considered a tool through which one can identify anomalous observations and develop instruments that facilitate invariant measurement. In contrast, proponents of the 2PL and 3PL models view of the role of a model as a representation of the empirical structure of data that should be modified to best represent observations. Thissen and Orlando (2001) summarize the perspective of 2PL and 3PL proponents, who advocate that "items are assumed to measure as they *do*, not as they should" (p. 90). Recognizing the philosophical controversy related to the classification of Rasch models as IRT models, this study will nonetheless refer to Rasch models as IRT models in order to facilitate the comparison between Rasch models and Mokken's (1971) nonparametric IRT models. Further, in order to match the notation used in IRT literature, the symbol θ and δ will be used to represent person and item locations on the latent variable, respectively, despite the use of β and δ for these values in traditional Rasch literature (see, for example: Wright & Stone, 1979).

Rasch Measurement Theory for Dichotomous Data

Rasch developed the Simple Logistic Model (SLM) – now referred to as the *Rasch model* – as part of a study to monitor the reading achievement of Danish students during the 1950s. While analyzing these achievement data, he recognized the useful properties of the SLM for describing the relationship between student achievement and item difficulty. First, the model overcomes challenges related to the raw score scale on which students and items are observed. Specifically, the raw score scale on which observations from an achievement test are often described does not facilitate a comparison between students and items because the scale units are not linear. In other words, the distance between points on a raw score scale may not have an equivalent interpretation across the range of the entire scale. More specifically, a one-point difference in the middle of the raw score scale does not have the same interpretation in terms of the latent variable as a one-point difference at the extreme ends of the raw score scale. Models based on Rasch measurement theory (Rasch, 1960/1980) overcome this challenge by describing student achievement and item difficulty on a linear logistic (i.e., logit) scale. When the Rasch model is applied, person and item raw scores undergo a nonlinear transformation that creates a scale onto which persons and items can be mapped that is more likely to have equal units. This *logit scale* describes students and items in terms of the log of the odds for a correct response at different locations on the latent variable. The raw-score transformation is used to estimate *person logits*, or θ estimates that represent person achievement, and *item logits*, or δ estimates that represent item difficulty in terms of the latent variable scale. A useful consequence of the linearity of the logit scale is that a difference in logit-scale locations of one unit between two students corresponds to a difference in the log-odds for a correct response of one, regardless of the particular item used for the comparison or the particular students being compared.

Although transforming total person and item scores to values on the logit scale facilitates comparisons between persons and items, the fact that the difference between person logits and item logits can range from $-\infty$ to $+\infty$ creates a problem when this difference is used to describe the probability for a correct response, which must range between zero and one. In order to relate the ability-difficulty difference to the probability for a correct response, the difference ($\theta_n - \delta_i$) can be raised to the natural constant, and the following ratio can be used to specify the probability for a correct response as a function of person ability and item difficulty:

$$\phi_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \qquad (8)$$

where

θ_n = the location of Person *n* on the construct (i.e., person ability),

δ_i = the location of Item *i* on the construct (i.e., item difficulty), and

ϕ_{ni1} = the probability for a correct response ($X = 1$) by Person *n* on Item *i.*

Equation 8 is called the Operating Characteristic Function (OCF) for the dichotomous Rasch model (Rasch, 1960/1980; Samejima, 1983). Essentially, an OCF describes the relationship between the probability for a response and the locations of persons and items on the latent variable. In the case of dichotomous items, the OCF for the Rasch model is the cumulative distribution function of the logistic distribution. Values from this function range between zero and one. As a result, the shape of the OCF is restricted to the shape of the ogives shown in Figure 7. As can be seen in the figure, the probability for a correct response (*y*-axis) increases as the difference between θ and δ (*x*-axis) becomes positive; in other words, the probability for a correct response increases as person achievement exceeds item difficulty.

For a single dichotomous item, the OCF is defined as an Item Response Function (IRF), or an Item Characteristic Curve (ICC); this study will use the term IRF to refer to OCFs for individual items. As will be shown later, OCFs are specified separately for each category of a polytomous item (Samejima, 1983). Figure 7 displays three Rasch IRFs.

An important characteristic of the Rasch model is that IRFs based on data that fit the model do not intersect. The implication of nonintersecting IRFs is that the order of item difficulties is probabilistically the same across all ability levels. In a parallel fashion, it is possible to define an OCF for persons, or Person Response Functions (PRFs). Because they are the person analog to Rasch IRFs, Rasch PRFs do not intersect. The result of nonintersecting PRFs is that a person who is located higher on the latent variable always has a higher chance of success on any item ($X = 1$) than a person who is located lower on the latent variable (Engelhard & Perkins, 2011). In other words, the order of person achievement is probabilistically the same across a set of items. Thus, a useful result of non-crossing IRFs and non-crossing PRFs is that invariant measurement is achieved: Person ability (θ) may be estimated without the influence of the effects of item difficulties (δ), and item difficulty (δ) may be estimated without the effects of person abilities (θ). Rasch (1960/1980, 1961) recognized invariance as a defining characteristic of his model that makes it compatible with the principles of measurement in the physical sciences (i.e., fundamental measurement). He summarized invariance in terms of persons and items as follows:

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison…. Symmetrically, a comparison between two individuals should be independent of which particular stimuli were instrumental for the comparison. (Rasch, 1961, pp. 331-332)

Mathematically, the Rasch model facilitates invariant measurement because the logarithm of the odds for a positive response divided by the probability of a negative response is $\theta - \delta$. Assuming independence of responses (discussed further below), the difference in probability for a correct response on Item *i* and a correct response on Item *j* can be estimated independently from any particular person. The logarithm of the ratio for success on Item *i* and success on Item *j* [P(X_{in} = 1 and X_{jn} = 0) / (X_{in} = 0 and X_{jn} = 1)] can be shown to equal $\delta_j - \delta_i$; this difference is independent from Student *n*. The invariance property establishes Rasch models as ideal-type models, which were discussed in Chapter One.

Rasch Model Requirements

The Rasch model is based on a set of underlying requirements that can be described in terms of the dimensionality, item independence, and functional form categories mentioned earlier in the general presentation of IRT. The third column of Table 5 provides a broad summary of the Rasch model requirements within these three categories. In the next section, the underlying requirements for Rasch models are described in terms of the three categories in Table 5. Because Rasch models are widely known in the educational measurement community, these underlying requirements will be discussed using less detail than in the description of the underlying assumptions for the Mokken models in Chapter Four. Additional details about Rasch models can be found in Bond and Fox (2007) and Engelhard (2013).

Dimensionality. The first major underlying requirement for Rasch models is related to dimensionality. In the context of IRT, dimensionality describes the number of latent variables that are modeled. Measurement models based on Rasch measurement theory require data to be unidimensional before measurement can be achieved. The psychological definition of unidimensionality is as follows:

- *Unidimensionality* – Item responses reflect evidence of a single latent variable.

Mathematically, unidimensionality can be examined by determining whether a single latent variable can be used to explain the structure of the response data. In contrast, multidimensional IRT models include multiple latent variables as explanatory variables for the structure of response data.

Item Independence. The second major underlying requirement for Rasch models is related to item independence. Specifically, Rasch models require that items demonstrate conditional independence; this concept is sometimes referred to as local independence. The psychological definition for conditional independence is as follows:

- *Conditional Independence* – Responses to an item are not influenced by responses to any other item, after controlling for the latent variable.

Essentially, conditional independence implies that the probability for a response is only determined by a person's location on the latent variable, and is not influenced by their response to any of the other items. Mathematically, the conditional independence requirement can be investigated by examining the covariance between two items, after controlling for a location on the latent variable (θ_n). If conditional independence is observed, this value will equal zero: $\mathrm{Cov}(X_i , X_j \mid \theta_n) = 0$.

Functional Form. The third category of model requirements is related to the functional form of an IRT model. The IRF that underlies the Rasch model implies that the probability for a response is a logistic function of the latent variable locations of persons and the difficulty of items. The IRF for the Rasch model was defined earlier as:

$$\phi_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad , \qquad (9)$$

where

ϕ_{ni1} = the probability for a correct response ($X = 1$) by Person n on Item i

θ_n = the location of Person *n* on the construct (i.e., person ability), and

δ_i = the location of Item *i* on the construct (i.e., item difficulty).

When adequate fit to the Rasch model is observed, the functional form of the IRF matches the shape of the IRFs shown in Figure 7. Because the Rasch model specifies the IRF using an algebraic formula (Equation 9), it is considered a parametric IRT model. Generally, the distinction between the terms *parametric* and *nonparametric* in statistics refers to the distribution, or shape, of a variable. In IRT, the distinction between parametric and nonparametric models refers to the properties of IRFs. Specifically, parametric IRT models restrict the shape of the IRF to a specific form. As is the case with the Rasch model, this form is usually the logistic or normal ogive. In the context of a rater-mediated educational assessment, the parametric form of the Rasch model suggests that the relationship between a student's location on the latent variable and the probability for an observed rating fits the shape of the logistic ogive.

The functional form of the Rasch model has two important consequences for measurement. First, there is a monotonic relationship between the latent variable and the probability for a correct response. In the case of dichotomous data, monotonicity in the latent variable implies that the probability that a person will correctly respond to an item [$P(X_i = 1)$] increases as their location on the latent variable (θ_n) increases. Stated another way, monotonicity in the latent variable implies that an increase in total scores (i.e., raw scores) corresponds to an increase in the estimated location on the latent variable. When data fit the Rasch model, the second important consequence of the functional form of the Rasch model is that the IRFs do not intersect. Because the Rasch model specifies the probability for a correct response as a function of the difference between person and item locations ($\theta_n - \delta_i$), the slope (i.e., discrimination) of each item is the same. This property of nonintersecting IRFs implies that estimates of item difficulty are invariant across the range of the latent variable, and that estimates of person achievement are invariant across a set of items.

Rasch Measurement Theory for Polytomous Ratings

In the last section, the dichotomous Rasch model was described in order to demonstrate the underlying principles and model requirements for models based on Rasch measurement theory. In addition to its usefulness for describing the relationship between persons, items, and latent variables in the case of dichotomous item response data, the Rasch model has also been formulated for use with rater-mediated assessments. Linacre (1989/1994) summarized the major motivation for the use of Rasch models in the context of rater-mediated assessment:

> For an examination in which judges rate examinees on test items, the ultimate goal of the judging process, from the viewpoint of an examining board, is not to determine some "true" rating for an examinee on each item, on which ideal judges would agree, but rather to estimate the examinee's latent ability level, of which each judge's rating is a manifestation. (p. 41)

This section describes Rasch measurement theory as a framework in which to consider rater-mediated assessments in general, and as a methodological tool to monitor the quality of ratings in particular. In order to addresses the third research question for this study (which asks: *How has Rasch measurement theory been used to examine the quality of ratings?*), Rasch models for ratings are considered in terms of their underlying requirements and in terms of their use in previous research to examine rater-mediated educational assessments.

Rasch Rater Model Requirements

Table 6 presents an adaptation of the IRT model requirements related to dimensionality, independence, and functional form for use with IRT models for raters. When raters are introduced to a measurement system, it is possible to conceptualize raters as assessment opportunities—similar to polytomous items on which students can receive scores in multiple categories. Thus, the requirements that define Rasch models for raters are based on interactions among students, raters, and latent variables. In the second column of Table 6, these requirements are summarized as they apply to Rasch models for rater-mediated assessments.

Dimensionality

Just as in the case of the dichotomous Rasch model, Rasch models for raters are based on the requirement of unidimensionality. Within the context of rater-mediated assessments, the Rasch model requirement of unidimensionality is stated in terms of ratings rather than items:

- *Rater Unidimensionality:* Ratings reflect evidence of a single latent variable.

Rater unidimensionality implies that ratings are not unduly influenced by construct-irrelevant variables, such as student characteristics (e.g., gender or handwriting), rater characteristics (e.g., rating or teaching experience), or characteristics of the assessment system (e.g., prompts or assessment consequences).

Independence

The next requirement is related to independence of assessment opportunities. The dichotomous Rasch model requires conditional independence for items. Considering raters as assessment opportunities, this requirement can be restated for use with rater-mediated assessments as follows:

- *Conditional Rater Independence*: The rating assigned to a student is not influenced by ratings assigned by other raters.

Functional Form

When Rasch models are applied to polytomous data, OCFs are estimated separately for each "step" in the rating scale, based on a conceptualization of the polytomous scores as a series of imaginary dichotomous (0/1) steps that reflect increasing amounts of the latent trait. For a rating scale with *k* categories, a student who receives a rating in Category *k* is viewed as earning a score of 1 (rather than 0) on the step between Category $k - 1$ and Category *k*. For example, if a student receives a score of '2' from a rater using a rating scale with three categories (0, 1, 2) it can be said that the score reflects a score of '1' (rather than '0') on the first category step (from 0 to 1), and a score of '1' on the second item step (from 1 to 2). Thus, there is a distinction between rating scale category steps and the final rater-assigned score of '2.' This view of polytomous ratings as a series of dichotomous steps is central to both the parametric and nonparametric IRT models employed in this study. In this study, these separate OCFs are called Category Response Functions (CRFs). For a rating scale that has *m* rating scale categories, there are $m - 1$ meaningful CRFs. In a parallel fashion to IRFs, CRFs describe the probability that a rater assigns a score in Category *k*, rather than Category $k – 1$, given a student's location on the latent variable. Variations on the Rasch model for rating scale data parameterize rating scale categories in different ways; these differences are discussed later in the chapter.

CRFs are based on the same functional form requirements as IRFs for the dichotomous Rasch model: monotonicity and nonintersection. In the context of a rater-mediated assessment, the monotonicity requirement for the category response function can be stated as:

- *Rater monotonicity:* The probability that a student will receive a higher rating increases as their location on a latent variable increases.

In other words, a student with a higher location on the latent variable is more likely to receive a higher rating than a student with a lower location on the latent variable. Next, the requirement of nonintersecting IRFs is extended to raters. In this study, the rater-analogue to IRFs is referred to as a Rater Response Function (RRF), and the nonintersection assumption is stated as follows:

- *Nonintersecting rater response functions*: Rater severity ordering is consistent across all levels of student achievement.

A result of nonintersecting RRFs is that person response functions also do not intersect, such that the ordering of persons in terms of the latent variable is consistent across a group of raters.

Using Rasch Measurement Theory to Examine the Quality of Ratings

Models based on Rasch measurement theory (Rasch, 1960/1980) have been used to explore a variety of issues related to rater-mediated assessments. Following the adaptation of the original Rasch model to polytomous rating scale data (Andrich, 1978; Linacre, 1989/1994; Masters, 1982), Rasch measurement theory has been presented as a framework for examining the quality of ratings in rater-mediated educational achievement tests (Eckes, 2011; Engelhard, 2002, 2013; Wolfe, 2009). Further, Rasch models have been presented as a method for detecting specific types of rater effects, such as rater errors (Eckes, 2005, 2008; Engelhard, 1994; Myford & Wolfe, 2003, 2004; Wolfe, 2004), rater accuracy (Engelhard, 1996; Wind & Engelhard, 2012), and the stability of rater severity over time (Congdon, & McQueen, 2000; Myford & Wolfe, 2009; Wolfe, Moulder, & Myford, 2001; Wolfe, Myford, Engelhard, & Manalo, 2007). In addition, Rasch models have been used to examine the impact of various rater training procedures (Knoch, Read, & von Randow, 2007; Weigle, 1998), score resolution procedures (Myford & Wolfe, 2002; Engelhard & Myford, 2003), and issues in the design of rating systems (Engelhard, 1997; Myford, Marr, & Linacre, 1996; Myford & Wolfe, 2000).

In this section, the two major formulations of the Rasch model for polytomous data are presented theoretically and mathematically: the Rating Scale model (Andrich, 1978) and the Partial Credit model (Masters, 1982; Wright & Masters, 1982). Then, the generalization of these models to the Many-Facet Rasch model (Linacre, 1989/1994) is presented, and previous applications of this model for examining rating quality are summarized. Finally, an illustrative data analysis with the Georgia writing data is used to demonstrate the application of models based on Rasch measurement theory as a methodological tool to examine the quality of ratings.

Rating Scale and Partial Credit Models for Ratings

Two formulations of the Rasch model for rating scale data are frequently used to examine properties of ratings: Andrich (1978) developed the Rating Scale (RS) formulation of the Rasch model, and Masters (1982) developed the Partial Credit (PC) formulation of the Rasch model. An important difference between these polytomous formulations of the Rasch model and the CTT approach to examining polytomous ratings is related to the parameterization of the rating scale categories. Specifically, models based on CTT assume that the distance between categories is equivalent across the range of the rating scale. For example, in the case of a four-point rating scale (0 = *Inadequate*; 1 = *Minimal;* 2 = *Good*; 3 = *Very good*), the CTT approach assumes that the distance between the first and second rating scale categories represents the same difference in terms of the latent variable as the distance between the third and fourth rating scale categories—a claim that may or may not be justified for a particular dataset. However, the CTT approach does not provide a method to empirically check this assumption. In contrast, Rasch models for rating data directly estimate the location of rating scale categories on the latent variable without the assumption of equidistant distances between each category. Both the PC and the RS formulations model a set of category response functions (CRFs) that describe the conditional probability that a student who has a particular location on the latent variable will receive a rating in a given category, rather than the category below. The major difference between the RS and PC models is related to the parameterization of the thresholds that distinguish rating scale categories on the logit scale. The RS model is stated as:

$$\ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \delta_i - \tau_k \quad , \qquad (10)$$

where

P_{nik} = the probability of Person n scoring k on Item i,

$P_{nik\text{-}1}$ = the probability of Person n scoring $k - 1$ on Item i,

θ_n = the location of Person *n* on the construct (i.e., person ability),

δ_i = the location of Item *i* on the construct (i.e., item difficulty), and

τ_k = the location on the construct where the probability of responding in

Category *k* and Category *k*−1 is equal across items.

When data are modeled using the RS model, category coefficient locations are fixed across items, indicated by the τ_k term. As a result, the distance on the latent variable between each pair of rating scale categories does not vary across tasks. Panel A of Figure 8 provides a graphical representation of the parameterization of category coefficients (τ) under the RS model.

In contrast, the PC formulation of the Rasch model (Masters, 1982; Wright & Masters, 1982) allows category coefficient locations to vary across items. The PC model is stated as:

$$\ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \delta_i - \tau_{ik}, \tag{11}$$

where

P_{nik} = the probability of Person *n* scoring *k* on Item *i*,

$P_{nik\text{-}1}$ = the probability of Person *n* scoring *k*−1 on Item *i*,

θ_n = the location of Person *n* on the construct (i.e., person ability),

δ_i = the location of Item *i* on the construct (i.e., item difficulty), and

τ_{ik} = the location on the construct where the probability for responding in

Category *k* and *k – 1* is equally probable for Item *i.*

The δ_{ik} term indicates that the category coefficient locations are estimated separately for each item. As a result, the PC model is essentially a test of the hypothesis of equidistant categories across items. As a result, it can be used as a diagnostic tool for comparing rating scale category use across a set of items. As illustrated in Panel B of Figure 8, this assumption may not be reflected in the observed use of rating scale categories. The PC model allows this hypothesis to be empirically investigated.

Many-Facet Rasch Model for Ratings

An adaptation of the Rasch model for polytomous data that is widely applied to rater-mediated assessments is the Many-Facet Rasch (MFR) model (Linacre, 1989/1994). The MFR model was originally developed as an approach to exploratory data analysis within the context of rater-mediated assessments. In contrast to the RS and PC models, which describe the probability for an observation as a function of person (θ) and rater (λ) locations on the latent variable, the MFR model can be specified as a generalization of either the RS or PC formulation of the Rasch model that incorporates additional explanatory variables, such as raters, tasks, and assessment occasions. These additional variables are called *facets.* Similar to a logistic regression model with fixed effects, the MFR model for ratings models observed ratings as the dependent variable with a single person parameter (θ) and additional researcher-specified facets as independent variables. When data fit the MFR model, invariant estimates of each of the independent variables on the logit scale can be obtained. Linacre (1989/1994) highlighted the usefulness of a measurement model that incorporates these facets of a rater-mediated assessment as a method for going beyond observed ratings in order to facilitate inferences about a latent variable. In his words:

> In order to supersede the local particularities of the judging situation, each judge must be treated as though he has a unique severity, each examinee as though he has a unique ability, each item as though it has a unique difficulty, and the rating scale as though it has one formulation applied identically by all the judges…. Thus each rating is considered to be the probabilistic result of only four interacting components: the ability of an examinee, the severity of a judge, the difficulty of an item, and the structure of the rating scale. With these assumptions, it is possible to obtain…an estimate of the ability of each examinee, freed from the level of severity of the particular judges who happened to rate the performance and also freed from the difficulty of the items and the arbitrary manner in which the categories of the rating scale has been defined. (p. 41)

Because it facilitates invariant measurement in situations involving multiple facets, the MFR model has been used to examine a variety of issues in educational assessments that require raters, judges, or panelists to assign polytomous ratings, including differential item and person functioning in large-scale writing assessments (Engelhard, 2009; Engelhard, Wind, Kobrin & Chajewski, in press), standard-setting procedures (Kaliski, Wind, Engelhard, Morgan, Reshetar, & Plake, 2013), second language assessments (McNamara, 1996, 2000), and problem-solving skills (Smith & Kulikowich, 2004). Particularly relevant to this study, the MFR model has been widely applied as a methodological tool for examining the quality of ratings in large-scale rater-mediated assessments. For example, Engelhard (2002, 2013) presents a set of criteria that can be used to evaluate the quality of rater-mediated assessments based on the MFR model.

Similar criteria have been proposed by Eckes (2011), Myford and Wolfe (2003, 2004), and Wolfe (2009).

Using the Georgia writing data, this section illustrates methods based on Rasch measurement theory (Rasch, 1960/1980) that can be used to evaluate rating quality in large-scale rater-mediated assessments. The presentation of Rasch-based rating quality indices is organized around two specifications of the MFR model. Model I is a MFR model for observed ratings that can be used to evaluate the degree to which a set of ratings meets the requirements for rater-invariant measurement. Model II is a MFR model that can be used to examine the concept of rater accuracy from the perspective of Rasch measurement theory.

Model I: Many-Facet Rasch Model for Rater Invariance

When the Many-Facet Rasch (MFR) model is applied to rating data, statistics and displays based on the model can be used to identify individual raters or groups of raters whose rating patterns suggest a lack of rater-invariant measurement (Eckes, 2011; Engelhard, 2013; Myford & Wolfe, 2003, 2004; Wolfe, 2009). Because it is an ideal-type model, indices and displays based on the MFR model can be used as evidence for rating quality in terms of the requirements for rater-invariant measurement. Engelhard (2002, 2013) extends the concept of rater errors and systematic biases to a set of statistics and displays based on the MFR model that can be summarized in three major categories: A) rater calibrations, B) model-data fit, and C) interactions. Table 7 displays indices of rating quality within these three categories using Rasch-based statistics that are calculated within the Facets computer program (Linacre, 2010). These statistics and displays can be viewed as indices of rater invariance.

The first model used in this study is a MFR model through which the major categories of statistics and displays shown in Table 7 can be used to examine rater invariance. Specifically, Model I is a three-facet formulation of the rating scale MFR Model (Wright & Masters, 1982; Linacre, 1989/1994):

$$\ln\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \theta_n - \lambda_i - \delta_j - \tau_k \quad , \tag{12}$$

where

P_{nijk} = probability of Student n receiving a rating of k by Rater i on Domain j,

$P_{nijk\text{-}1}$ = probability of Student n being rated $k – 1$ 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_k = difficulty of Category k relative to Category $k - 1$.

The dependent variable in this model is the log of the odds that a student receives a rating in Category k, rather than in Category $k – 1$, given their location on the latent variable, the severity of the rater, and the difficulty of the domain. A major benefit of the MFR model is that estimates for each facet are calculated on the log-odds, or logit scale, which represents the latent variable (in the case of the example dataset, writing achievement). Because estimates for each facet are described on the same scale, it is possible to compare latent-variable locations across facets. Logit-scale locations for students and raters under Model I describe writing achievement and rater severity for each student and rater, respectively. Similarly, domain calibration on the latent variable describes the judged difficulty of each of the four analytic rubric domains across the set of raters. The tau term in the model (τ_k) does not represent a facet; instead, this term represents the category coefficients for the rating scales used to score the essays. In addition, interaction terms can be added to the model to examine differences among rater interpretation of domain difficulty. Standard errors for each of these facets describe the precision of the measurement on the logit scale.

A. Rater Calibrations

As shown in Table 7, the first category of rating quality indices based on Model I is rater calibrations. First, rater leniency and severity can be examined by comparing the estimates of rater locations on the logit scale (λ) obtained from Model I. Because the logit scale serves as an operational definition of the latent variable, logit-scale locations for individual raters describe their interpretation of the level of achievement required to receive ratings in a particular category. Using the example of a writing assessment, raters with high locations on the logit scale require “more” writing achievement for higher ratings, and thus are considered severe; raters with low locations on the logit scale require “less” writing achievement, and thus are considered lenient. The logit-scale location for each rater summarizes their overall leniency and severity across domains and rating scale categories. Standard errors are calculated for each rater whose values describe the precision with which each rater is calibrated on the logit scale.

Results: Georgia writing data. Most software programs for MFR model analyses summarize logit-scale locations using a visual display called a *variable map*. Figure 9 is a variable map that summarizes the results from Model I for the Georgia writing data. Specifically, Figure 9 displays student writing achievement measures, rater severity calibrations, domain difficulty locations, and the rating scale categories on a common linear logistic scale. This visual display represents the overall shared understanding of student writing achievement, domain difficulty, and rating-scale categories for the Georgia High School Writing Test among the sample of raters. The first column is the logit scale that represents writing achievement, which is the latent variable examined in this study. The next three columns display the logit-scale locations for the three facets in Model I: Students, Raters, and Domains. In order to provide a frame of reference for interpreting the logit-scale locations of the three facets, raters and domains are centered at zero (mean set to zero), and only the average location of the student facet is allowed to vary. The second column displays the student locations on the latent variable (*n* = 365). As can be seen in the variable map, there is a wide range in student locations on the latent variable, which suggests that the group of raters detected differences among these students in terms of writing achievement. Students who are located higher on the logit scale received higher ratings, and students who are located lower on the logit scale received lower ratings. Examination of the Rater column reveals that there is not much variation among the locations of the raters on the logit scale compared to the spread of students. Raters who are located higher on the logit scale were more severe; i.e., they assigned lower ratings more often. Raters who are located lower on the logit scale are less severe; i.e., they assigned higher ratings more often. Finally, the location of the domains on the logit scale reflects the judged difficulty of the style, organization, conventions, and sentence formation domains. Domains that are located higher on the logit scale are associated with more severe (i.e., lower) ratings, and domains that are located lower on the logit scale are associated with less severe (i.e., higher) ratings. Table 8 summarizes these locations in terms of the mean location for each facet and the standard deviation for the spread of elements within each facet.

Logit-scale locations for raters based on Model I describe the overall severity and leniency of individual raters. These estimates can be used to identify raters who are systematically higher or lower than the rest of a group. Further, standard errors for rater severity estimates can be used to describe the precision with which individual rater severity measures are calibrated. Table 9 summarizes the individual rater calibrations for the Georgia writing data. As can be seen in the table, rater severity measures range from −0.56 logits (*SE* = 0.05) for Rater 15, who is lenient, to 0.73 logits (*SE* = 0.05) for Rater 9, who is severe.

After estimates of the main-effect parameters are calculated, several statistics can be examined to identify further characteristics of the data related to rating quality. This study focuses on separation statistics and model-data fit statistics. First, the *reliability of separation* statistic based on Rasch models is an index of how well individual elements within a facet can be differentiated from one another, such as individual persons or raters. The reliability of separation statistics for persons is comparable to Cronbach's coefficient alpha and KR20 because it reflects an estimate of true score to observed score variance. For the other facets, the reliability of separation statistic describes the spread or differences between elements within a facet, such as differences in rater severity. The statistic is calculated using the standard deviation (*SD*) and Mean Square Error (*MSE*) as follows:

$$Rel = \frac{(SD^2 - MSE)}{SD^2} \quad , \qquad (13)$$

where SD^2 is the observed variance of elements within a facet in logits and *MSE* is the mean square calibration error. *MSE* is estimated as the average value of calibration error variances (squares of the standard errors) for each element within a facet. Andrich (1982) provides a detailed derivation of this reliability of separation index. In addition to the reliability of separation statistic, a chi square statistic (χ^2) is calculated to test the null hypothesis that the differences between elements within a facet are not significantly different from zero. The chi square test provides a method for determining whether a group of raters can be considered exchangeable; significant differences among individual raters suggest that raters are not exchangeable.

Results: Georgia writing data. In the example dataset, the overall differences between elements within the Student (θ), Rater (λ), and Domain (δ) facets are significant ($p < 0.05$), and the reliability of separation for students (equivalent to coefficient alpha) is quite high ($Rel_\theta = 0.99$). The reliability of separation statistic for students is interpreted in the same manner as Cronbach's alpha. For the Rater and Domain facets, the reliability of separation statistic describes the spread, or differences, between elements within a facet. High reliability of separation statistics for raters (Rel_λ= 0.98) and domains ($Rel_\delta > 0.99$) suggest that there are significant differences among the raters in terms of severity and among the difficulty estimates of the four domains examined in this study.

B. Model-Data Fit

Next, indices of rating quality based on the MFR model include a set of statistics and displays related to model-data fit. Because the MFR model meets the requirements for invariant measurement, departures from model expectations, or residuals, are of interest for describing rating quality. Thus, the Rasch-based approach to examining rating quality focuses on an examination of standardized residuals, summaries of residuals in the form of model-data fit statistics, and graphical displays of residuals.

First, residuals are calculated that compare observed ratings to the expected ratings based on the MFR model:

$$Y_{nij} = X_{nij} - P_{nij} \tag{14}$$

where

X_{nij} = observed rating, and

P_{nij} = expected rating probability based on the MFR model (Equation 12).

The residuals are standardized as

$$Z_{nij} = \frac{Y_{nij}}{\sqrt{Q_{nij}}} \quad , \tag{15}$$

with Q_{nij} defined as the response variance, or statistical information, for each item:

$$Q_{nij} = P_{nij}(1 - P_{nij}) \quad . \tag{16}$$

These residuals are summarized as fit statistics that describe the degree to which adherence to the requirements for invariant measurement is observed in a set of data. Fit statistics are used within the Rasch-based approach in order to examine the degree to which adherence to the requirements for invariant measurement is observed in a set of data.

The approach to model-data fit analysis within Rasch measurement theory typically focuses on fit statistics that summarize residuals, or differences between model expectations and empirical observations. In this study, model-data fit is explored in terms of two fit statistics: Infit and Outfit *MSE* statistics. These statistics are routinely used in

Rasch analyses for facets related to persons and items (e.g., Engelhard, 2002; Wolfe, 2009). In this study, the item-related fit statistics are applied to raters.

The Outfit *MSE* statistic is calculated by summing standardized residual variance across facets. Because it is unweighted, the Outfit *MSE* statistic is useful because it is particularly sensitive to "outliers," or extreme unexpected observations. The person Outfit *MSE* (U_n) statistic is calculated as follows:

$$U_n = \frac{\sum_{i}^{L} Z_{ni}^2}{L} \quad , \tag{17}$$

where Z^2_{ni} represents standardized score residuals and *L* is the number of raters. Similarly, the Outfit MSE statistic for raters (U_i) is calculated as:

$$U_i = \frac{\sum_{n}^{N} Z_{ni}^2}{N} \tag{18}$$

where *N* is the number of persons.

Infit *MSE* statistics are also useful for evaluating model-data fit. However, they are less sensitive to outlying data because the residuals are weighted by the variance of an individual facet, which reduces the impact of unexpected observations. Similar to Outfit *MSE*, Infit *MSE* can be calculated for person- and rater-related facets. The Infit *MSE* statistic for persons (U_n) is calculated as:

$$U_n = \frac{\sum_{i=1}^{L} Y_{ni}^2}{\sum_{i=1}^{L} Q_{ni}} \quad , \tag{19}$$

and the Infit *MSE* statistic for raters (V_i) is calculated as:

$$V_i = \frac{\sum_{n}^{N} Y_{ni}^{2}}{\sum_{n}^{N} Q_{ni}} \quad , \qquad (20)$$

where Y^2_{ni} represents score residuals for raters and Q_{ni} is an estimate of response variance (statistical information).

Although limitations of Rasch fit statistics have been noted in previous research (e.g., Karabatsos, 2000; Smith, Schumacker, and Bush, 2000), useful applications of Infit and Outfit *MSE* statistics have been demonstrated (Engelhard, 1994; Linacre, 1989/1994). Because the exact sampling distribution for these fit statistics is not known (Wright and Masters, 1982; Engelhard, 2013), various rules of thumb have been proposed for interpreting their values as they apply to specific types of facets, such as raters and students. Engelhard (2009) describes an acceptable range of Infit and Outfit *MSE* statistics of about 0.80 to 1.20. Values that are lower than about 0.80 suggest possible dependencies among ratings, and values that are higher than about 1.20 suggest haphazard ratings; extreme values in both directions warrant further investigation. Engelhard (1994) describes the application of Infit and Outfit MSE statistics to rater-mediated writing assessment as a method for identifying *response sets*, which are a category of rater errors related to idiosyncratic use of a rating scale. Specifically, low values of fit statistics may suggest that raters are not making full use of a rating scale, or that there are dependencies across the ratings assigned to a group of students. On the other hand, high values of fit statistics may suggest haphazard ratings, or erratic use of the rating scale.

Results: Georgia writing data. For the Georgia writing data, Table 8 summarizes fit statistics for each facet in Model I. Overall, fit statistics for the Student, Rater, and Domain facets suggest that the MFR model is functioning as intended for these rating data. Table 9 provides Infit and Outfit *MSE* statistics for each of the 20 operational raters. As can be seen in Table 9, the spread of Infit and Outfit *MSE* statistics for ratings indicated acceptable fit to the model, with the lowest Outfit *MSE* value observed for Rater 9 (Outfit *MSE* = 0.73) and the highest value observed for Rater 18 (Outfit *MSE* = 1.27). The lowest Infit *MSE* value was observed for Rater 9 (Infit *MSE* = 0.76) and the highest value for Rater 21 (Infit *MSE* = 1.26).

In order to further examine model-data fit for these raters, Rasch-based analyses of rating quality include the use of visual displays of standardized residuals for raters whose fit statistics may suggest idiosyncratic rating patterns. Focusing on Outfit *MSE* statistics because they are sensitive to extreme departures from model expectations, Figure 10 illustrates standardized residuals for raters with low, expected, and high values of Outfit *MSE* statistics. Standardized residuals are plotted for the three selected raters across the group of 365 students (*x*-axis), and residual values greater than |2.00| (y-axis) indicate statistically significant differences between observed and expected rater severity for individual students, based on the overall locations of raters and students on the logit scale. As can be seen in the figure, statistically significant residuals are most frequent for Rater 18, whose ratings are “noisy,” and least frequent for Rater 17, whose ratings are “muted.”

C. Interactions

The third category of rating quality indices based on Model I is related to interactions between rater severity and other facets in the MFR model. In the context of a rater-mediated assessment, interactions provide a method for identifying whether or not rater severity is invariant across facets of interest. Interactions between rater severity and *internal facets*—facets related to an aspect of the assessment system, such as domains—can be used to identify raters whose interpretation of a rubric may be different from the rest of a rater group. Along the same lines, interactions between rater severity and *external facets*—facets related to something beyond the assessment system, such as student gender—can be used to identify *differential rater functioning*. Similar to differential item functioning, differential rater functioning occurs when raters are systematically more or less severe when rating students with particular construct-irrelevant characteristics (Engelhard, 1994, 2002; Wolfe, Moulder, & Myford, 2001).

Interaction terms can be added to MFR Models in order to examine whether rater interpretation of the construct is consistent across facets. Tests for significant interactions based on the MFR model test the null hypothesis that there is no significant interaction between the logit-scale location for particular facets. Within the context of a rater-mediated writing assessment, interactions between rater severity and the writing domains provide information about the quality of ratings assigned within each domain. First, an interaction term between the two facets of interest is added to the model:

$$\ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = (\theta_n - \lambda_i - \delta_j - \mu_m - \tau_k) - \lambda_i\delta_j\,, \tag{21}$$

where $\lambda_i\delta_j$ is the interaction between rater severity and domain difficulty.

The Facets computer program (Linacre, 2010) can be used to compute an overall fixed chi square (χ^2) test for the significance of a set of interactions. This test statistic is used to confirm or disconfirm the null hypothesis that the overall set of interactions is not significantly different from zero, after allowing for measurement error. In other words, the significance test for this statistic answers the question: *Is the overall set of interactions between these two facets significantly different from zero?* A significant value for this test suggests that interactions between the two specified facets are significant at an overall level. For the interaction between raters and domains, the omnibus test examines the null hypothesis that the interactions between individual raters and the judged difficulty of each domain are not significant. In addition to the overall test for significant interactions, it is possible to examine individual interactions between two facets. These individual terms provide information about the direction and magnitude of each interaction, and they are useful for identifying patterns in data that can be used to inform the interpretation of measurement outcomes. Visual displays that illustrate these interaction terms are particularly useful for this purpose. Individual interaction terms are calibrated on the logit scale, and their precision is described using standard errors. As a result, it is possible to evaluate the significance of the interactions using *t*-tests. A type of standardized effect size, these test statistics are primarily used as descriptors of patterns within rating data that may signal rater effects of substantive interest. As a result, significance tests are not of primary importance. However, a general practice in the interpretation of these statistics is to use an absolute value of 2.00 to identify interactions that may warrant further investigation (Engelhard, 2009). In this study, interactions suggest that rater severity is higher or lower than expected for a domain, based on its expected location on the logit scale (across the group of raters from Model I).

Results: Georgia writing data. Figure 11 illustrates interactions between rater severity and writing domain difficulty. For clarity, interactions are displayed separately within each of the four domains. Raters are listed along the *x*-axis, and the test statistic for the null hypothesis that there is no interaction effect between rater severity and domain difficulty is plotted along the *y*-axis. Lines that mark the critical values of positive and negative two are used to highlight significant *t*-statistics. Values of the *t*-statistic greater than an absolute value of two ($t > |2|$) suggest that rater severity in a domain is higher or lower than expected, respectively, based on its overall difficulty measure. Inspection of Figure 11 reveals that there are significant interactions between rater severity and domain difficulty within each of the four domains. Significant interactions between these two facets suggest that rater severity may not be invariant across the rubric domains on the Georgia High School Writing Test. In other words, significant interactions suggest that raters may be systematically more or less severe when assigning ratings within a particular domain than would be expected if their ratings matched the expectations of the ideal-type model. Significant interactions between these two facets were most frequently observed in the Conventions domain.

The second type of interaction analysis involves the examination of interactions between rater severity and external facets. Engelhard (2007) illustrates these interaction analyses as a method for detecting differential rater functioning (DRF) across student subgroups. In order to explore this type of interaction within the Georgia writing data, a student gender facet is added to Model I:

$$\ln\left[\frac{P_{nijmk}}{P_{nijmk-1}}\right] = \theta_n - \lambda_i - \delta_j - \mu_m - \tau_k , \qquad (22)$$

where μ_m represents the average writing achievement within gender subgroups (male and female students), and the other terms are defined as before. Then, an interaction term can be added between rater severity and student gender ($\lambda_i\mu_m$):

$$\ln\left[\frac{P_{nijmk}}{P_{nijmk-1}}\right] = (\theta_n - \lambda_i - \delta_j - \mu_m - \tau_k) - \lambda_i\mu_m . \quad (23)$$

Interactions between rater severity and student gender are examined in the same manner as described above for the interaction between rater severity and writing domains.

Results: Georgia writing data. Figure 12 illustrates interactions between rater severity and student gender within the example dataset. For clarity, interactions are displayed separately for students in the female and male subgroups. Inspection of Figure 12 reveals that significant interactions between rater severity and student gender. The finding of significant interactions between these two facets suggests that rater severity is not invariant across student gender subgroups on the Georgia High School Writing Test.

Model II: Many-Facet Rasch Model for Rater Accuracy

In addition to indices of rater invariance, the MFR model can also be used to evaluate rater accuracy in a direct way when expert ratings are available. Engelhard (1996, 2013) describes an application of the MFR model to describe rater accuracy based on a match between operational and expert ratings. The first step in modeling rater accuracy within a Rasch framework is to compute dichotomous accuracy scores for each observed rating. An accuracy score of zero is assigned when there is a discrepancy between the rating assigned by an operational rater and an expert rater. In contrast, an accuracy score of one reflects a perfect match between an operational and expert rater. Using this dichotomous scoring scheme, rater accuracy can be modeled as a latent variable on which raters, essays, and domains are calibrated. Table 10 summarizes and extends Engelhard's (1996, 2013) indices of rater accuracy based on the MFR model. These indicators of rater accuracy are illustrated below with the Georgia writing data.

Model II is a two-facet formulation of the MFR Model, and it is used to illustrate Rasch-based indices of rater accuracy using the Georgia writing data. The example dataset includes scores from a validity committee for each of the 365 essays. In order to conduct accuracy analyses, a second data set of dichotomous accuracy scores was created based on a comparison of operational ratings to the expert ratings from the validity committee. An accuracy score of zero was assigned in the case of a discrepancy between the rating assigned by an operational rater and an expert rater. In contrast, an accuracy score of one reflects a perfect match between an operational and expert rater. Using this dichotomous scoring scheme, rater accuracy can be modeled as a latent variable on which raters, essays, and domains are calibrated. Although a polytomous scheme for rater accuracy that describes the magnitude of the difference between operational and expert raters is also possible, this study will apply the dichotomous conceptualization of rater accuracy due to its use in previous research (Engelhard, 1996, 2013). These dichotomous accuracy scores were applied to Model II:

$$\ln\left[\frac{P_{nij(X=1)}}{P_{nij(X=0)}}\right] = \beta_n - \delta_i - \lambda_j \quad , \tag{24}$$

where

$P_{nij(\mathrm{X}=1)}$ = probability of Benchmark Essay *n* scored by Rater *j* within Domain *i* being rated accurately (X = 1),

$P_{nij(\mathrm{X}=0)}$= probability of Benchmark Essay *n* scored by Rater *j* within Domain *i* being rated inaccurately (X = 0),

β_n = difficulty of providing an accurate rating on Benchmark Essay *n*,

λ_j = accuracy "ability" of Rater *j*,

δ_i = difficulty for accurate rating of Domain *i*.

The object of measurement in Model II is the rater, and the construct is rater accuracy. Similar to Model I, statistics and displays based on Model II can be used to describe rater accuracy in terms of three major categories: A) rater accuracy calibrations; B) model-data fit; and C) rater accuracy interactions. These rating quality indices parallel the Rasch-based statistics and displays for Model I, and rating quality indices based on accuracy scores are calculated in the same manner as in Model I. Results from Model II analyses with the Georgia writing data are presented below in order to illustrate the use of this model in the context of large-scale rater-mediated assessment.

A. Rater Accuracy Calibrations

The first category of rater accuracy indices based on Model II is related to the calibrations of raters on a linear scale that is used to represent rater accuracy. Indices within this category include estimates of rater locations on the construct and separation statistics that describe differences in rater severity across a group of raters.

Results: Georgia writing data. Figure 13 summarizes findings from the Facets (Linacre, 2010) analysis for rater accuracy using Model II. In the context of accuracy ratings, the variable map represents rater accuracy in terms of the difficulty for raters to provide accurate ratings on each student essay, individual rater accuracy measures, and the difficulty for raters to provide accurate ratings within each of the writing domains. The first column is the logit scale that represents rater accuracy. The second column displays the location of each essay on the accuracy scale. The location of the essays on the logit scale represents the difficulty for operational raters to assign accurate scores. As can be seen in the variable map, there is a wide range in location of the 365 essays on the logit scale, which suggests that there are differences among the essays in terms of difficulty for the operational raters to assign accurate scores. Essays that are located higher on the logit scale received inaccurate ratings more often, and essays that are located lower on the logit scale received accurate ratings more often. Examination of the Rater column reveals that there is some variation among the locations of the raters on the logit scale. Raters who are located higher on the logit scale were more accurate; i.e., their ratings matched expert ratings often. Raters who are located lower on the logit scale are less accurate; i.e., they assigned ratings that matched the expert less often. Finally, the location of the domains on the logit scale reflects the difficulty for raters to assign accurate ratings within each of the domains on the Georgia High School Writing Test. Domains that are located higher on the logit scale are associated with more accurate ratings, and domains that are located lower on the logit scale are associated with less accurate ratings. Table 11 provides a summary of the logit scale locations for the Rater, Student, and Domain facets in Model II; these locations correspond to the variable map shown in Figure 13.

Table 12 summarizes the calibration of the rater facet under Model II. 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. As can be seen in the table, the overall differences between rater accuracy calibrations (β), essay difficulty for accuracy (δ), and domain difficulty for accuracy (α) are significant ($p < 0.05$), with high reliabilities of separation

($Rel_\beta = 0.97$; $Rel_\delta = 0.83$; $Rel_\alpha = 0.84$). Although the difference between domain locations was statistically significant, examination of the logit-scale locations for the domain facet suggests that these differences may not be substantively meaningful.

B. Model-data Fit for Accuracy

Indices of model-data fit for Model II have a slightly different interpretation than the *MSE* statistics based on Model I. When accuracy scores are modeled, standardized residuals and Infit and Outfit statistics describe the match between model expectations and observed accuracy scores for individual raters. Raters whose fit statistics are noisy ($\geq$ 1.20) or muted ($\leq$ .80) display more or less variation, respectively, in their ability to assign accurate ratings to a set of performances than expected by the dichotomous Rasch model for rater accuracy. Raters with *MSE* statistics around 1.00 match model expectations. Similarly, fit statistics for Student and Domain facets describe the match between model expectations and rater accuracy for specific students and domains.

Results: Georgia writing data. As can be seen in Table 11, acceptable fit to the model is evident for each of the facets in Model II, with mean Infit and Outfit *MSE* statistics around 1.00, and standard deviations around 0.20. Acceptable model-data fit suggests that the Rasch model is functioning as intended for these dichotomous accuracy data. For the rater facet, Table 12 indicates that the spread of Infit and Outfit *MSE* statistics for individual raters suggest 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* = 1.05). Because all of the *MSE* fit statistics for this group of raters are between 0.80 and 1.20, further examination with standardized residual plots is not necessary.

C. Rater Accuracy Interactions

The third category of direct indices of rater accuracy based on the MFR model includes the use of rater accuracy interactions. These interactions describe the degree to which rater accuracy is invariant over internal and external facets, and they are calculated and interpreted in a similar fashion as was described for Model I. In order to illustrate these rating quality indices, two interactions are explored with the Georgia writing data: the interaction between rater accuracy and domains, and the interaction between rater accuracy and student gender.

Results: Georgia writing data. Figure 14 illustrates interactions between the rater accuracy and the domain facets in Model II (βδ). For clarity, interactions are displayed separately for each domain. Inspection of Figure 14 reveals that, in general, rater accuracy does not appear to interact significantly with domains. In order to examine the interaction between rater accuracy and student gender, a gender facet was added to Model II that represents the average rater accuracy for female and male students (μ). Then, an interaction term was added between rater accuracy and student gender (βμ). Figure 15 illustrates results from this interaction analysis. For clarity, interactions are displayed separately for male and female students. Inspection of Figure 15 reveals that, in general, there are no significant interactions between rater accuracy and student gender; in other words, rater accuracy is invariant over student gender subgroups.

Summary

The Rasch measurement theory approach to evaluating rating quality describes individual raters in terms of their unique locations on the latent variable, and compares observed ratings to patterns expected by the ideal-type model. In terms of the lens model presented in Chapter One (Figure 2), the Rasch perspective for describing the match between θ_P and θ_R is based on evidence of fit to an ideal-type model. As Linacre (1989/1994) observed, this perspective of rating quality focuses on the useful properties of the Rasch model as a criterion for evaluating rating quality, while summarizing aberrant observations (e.g., rater errors and systematic biases) through indices of model-data fit:

> The more that incidental aspects of behavior are in evidence in the ratings, the more uncertainty there is in the estimate of each examinee's ability, and the less confidence there is that the aim of the judging process has been realized in the judge's ratings. Thus accurate measurement depends not on finding the one "ideal" judge but in discerning the intentions of the actual judges though the way in which they have replicated their behavior in all the ratings each has made (p. 41).

Consequently, there is an important philosophical difference between the view of rating quality based on Rasch measurement theory and the traditional approach that focuses on comparing individual raters to other raters in a group in terms of agreement, error, and accuracy (see Chapter Two). Whereas the traditional approach focuses on replications of "ideal" raters who agree with one another, demonstrate a lack of errors and systematic biases, and match the ratings assigned by experts, Rasch models for ratings are probabilistic, such that variance in ratings is necessary in order to construct measures from ratings that describe students, raters, and other tasks in terms of a latent variable.

Chapter Four: Illustration of Modern Rating Quality Indices based on Mokken Scaling

This chapter continues the illustration of modern rating quality indices from Chapter Three by presenting a second set of rating quality indices based on Mokken's (1971) theory and procedure for scale analysis. Mokken's approach to scaling includes a set of models that can be described using the IRT framework that was presented in Chapter Three. Following the structure of Chapter Three, this chapter contains two major sections: First, an overview of Mokken scaling is provided, and the dichotomous formulations of Mokken's nonparametric models are described theoretically and mathematically; Then, the polytomous versions of these models are presented and extended for use as indicators of rating quality. As in the previous chapter, data from the Georgia High School Writing Test are used to illustrate rating quality indicators based on Mokken scale analysis.

What is Mokken Scaling?

The nonparametric models used in this study are based on Mokken scale analysis (Mokken, 1971). Mokken's *A Theory and Procedure for Scale Analysis* presents an approach to social science measurement that combines key ideas from Guttman scaling (Guttman, 1950), and probabilistic item response models that facilitate invariant measurement, such as the Rasch model (Rasch, 1960/1980). Using examples from political research, Mokken offered a systematic method for developing and validating scales to measure latent variables. Similar to Rasch measurement theory, Mokken's nonparametric models describe the probability for an observed response in terms of person and item locations on a latent variable. In contrast to Rasch models, the Item Response Functions (IRFs) that underlie Mokken's models are not required to conform to the shape of the logistic ogive—as a result, these models are described as nonparametric item response theory (nonparametric IRT) models (Sijtsma & Molenaar, 2002). Nonparametric IRT models are considered less restrictive than parametric IRT models because of the lack of restrictions imposed on the shape of the IRFs.

Specifically, Mokken's (1971) nonparametric IRT models describe the probability for a correct or positive response in terms of a function that is only governed by order restrictions, such that the IRF ($P_i| \theta$) is a nondecreasing function of the latent variable (θ). For example, for persons *A* and *B* who can be ordered in terms of the latent variable such that ($\theta_A < \theta_B$), the only restriction on the IRF is the assumption that the following relationship holds:

$$(P_{i=1}|\theta_A) \leq (P_{i=1}|\theta_B) \quad , \tag{25}$$

where

($P_{i=1}| \theta_A$) = probability that a person at ability level *A* provides a correct response to Item *i*, and

($P_{i=1}| \theta_B$) = probability that a person at ability level *B* provides a correct response to Item *i*.

As a result, the IRFs associated with nonparametric IRT may take on a variety of shapes. Three nonparametric IRFs that meet the ordering assumption in Equation 25 are illustrated in Figure 16. Several things are important to note about the shape of nonparametric response functions. First, the IRFs do not need to match the *s*-shape of the logistic or normal ogive, as is the case with most parametric IRT models. Second, nonparametric IRFs can have several points of inflection, and they may be constant over several ranges of the latent variable. Finally, nonparametric IRFs do not need to cover the full range of the *y*-axis; that is, the lower asymptote may be higher than zero, and the upper asymptote can be lower than one. Molenaar (1982, 1997) extended Mokken's (1971) method for scale analysis for use with polytomous data. Similar to parametric IRFs, nonparametric IRT models for polytomous data are based on a conceptualization of rating scale categories as a series of dichotomous "steps," such that $m - 1$ separate response functions are specified for a rating scale that has m unique categories. Within the framework of nonparametric IRT, Category Response Functions (CRFs) are defined as the cumulative probability for a rating within a category. These cumulative category probabilities are constrained by the same order restriction as the dichotomous response functions (Equation 25).

Motivation for Nonparametric IRT

Essentially, the use of nonparametric IRT models for item response data is motivated by a lack of confidence in the tenability of the functional form restrictions that underlie their parametric counterparts. Specifically, these reservations are related to the prerequisite that the relationship between person locations on the latent variable and the probability for a correct response fit the shape of the logistic or normal ogive. Sijtsma and Meijer (2007) described several useful results of the limited restrictions that characterize IRFs in nonparametric IRT. Contrasting the nonparametric approach with parametric IRT, they observed that parametric IRT models can be viewed as more parsimonious, because the relationship between items and persons is governed by a limited number of parameters. However, they emphasize the diagnostic value of placing fewer restrictions on the shape of the IRF. In their words,

> Estimating a [parametric] model is like stretching a grid over the data which is flexible only in some directions but not in others, and then summarizing the curvature and location of the grid as far as its flexibility allows it. … Obviously, peculiarities in the data may not be caught by logistic or other parametric curves but, if properly revealed, could help the researcher to make useful decisions about the items. (p. 726)

As an example, they noted that a nonparametric IRF may be useful for diagnosing areas on the latent variable where the monotonicity assumption does not hold for an item of interest. This example highlights the utility of nonparametric IRT for identifying items that satisfy or do not satisfy the criteria for a particular measurement purpose. Along the same lines, Mokken (1971) summarized the rationale for a nonparametric approach to item response modeling:

> In vast areas of social research the application of parametric models may often be too far fetched. Their application presupposes a relatively deep insight into the structure of the variable to be measured and the properties of the items by which it can be measured. For these reasons it seems legitimate to try to find starting points for scaling models which do not rely too heavily upon specific parametric assumptions, as these lead to procedures of inference and estimation that are too pretentious and intricate for the level of information and the precision that can be claimed for the data used in actual measurement. (p. 173)

Based on these reservations, nonparametric IRT is often proposed as a starting point that can be used to explore the tenability of parametric model assumptions. Sijtsma and Meijer (2007) pointed out that the difference between a parametric and nonparametric approach to IRT can be viewed in terms of the difference between exploratory and confirmatory data analysis. From this perspective, nonparametric methods provide a broad framework for understanding parametric IRT, and are essentially an "exploratory toolkit" in which the basic requirements for parametric models can be explored. In contrast, parametric IRT models are confirmatory in the sense that they are used to examine the degree to which empirical data match the form of pre-specified IRFs.

Mokken Model Assumptions

Mokken (1971) proposed two nonparametric IRT models that describe relationships among items, persons, and latent variables: the Monotone Homogeneity (MH) model and the Double Monotonicity (DM) model. Returning to the three major categories of IRT model assumptions discussed in Chapter Three (see Table 5), the MH and DM models are based on assumptions related to dimensionality, item independence, and the functional form that relates persons, items, and latent variables. The fourth column of Table 5 provides a broad summary of the underlying assumptions for Mokken's (1971) models within these three categories. Inspection of Table 5 reveals that the MH and DM models share the assumptions of unidimensionality and conditional independence with the Rasch family of models. As given earlier, these assumptions are as follows:

- *Unidimensionality*: Item responses reflect evidence of a single latent variable.

- *Conditional Independence*: Responses to an item are not influenced by responses to any other item, after controlling for the latent variable.

The distinguishing feature of Mokken's (1971) models from the Rasch family of models is related to the functional form assumption. Further, differences related to the functional form assumption distinguish the MH model from the DM model. Methods for investigating model-data fit within the framework of nonparametric IRT focus on checking observable properties of the functional form assumptions of the MH and DM models. In this section, the functional form assumptions that underlie the MH and DM models are described. Then, methods for evaluating these assumptions and extensions of these models for rater-mediated assessments are described and illustrated using the Georgia writing data.

Mokken Scaling for Dichotomous Data

Mokken (1971) originally proposed his method for scale analysis using dichotomous data. In this section, the Monotone Homogeneity and Double Monotonicity models are described in terms of their original dichotomous formulations. Then, polytomous versions of these models are introduced and extended for use with rater-assigned scores.

Monotone Homogeneity

Mokken's (1971) Monotone Homogeneity (MH) Model is a nonparametric IRT model based on the assumptions of unidimensionality (homogeneity) and conditional independence. In terms of functional form, the MH model does not require the probability function for observed responses to conform to a specific shape, as long as the item response functions are monotonically increasing in the latent variable (monotonicity). In the context of item-response data, monotonicity in the latent variable suggests that the probability that a student correctly responds to an item ($X_i = 1$) increases as their ability level increases. In other words, students with higher ability levels should have a higher chance for success on items than students with lower ability levels. Because nonparametric IRT procedures do not allow for the direct estimation of student locations on the latent variable (θ estimates in parametric IRT), an approximation of the student ability parameter is obtained by calculating total scores (X_+) for each student across an entire set of items. Mokken (1971) demonstrated that an ordering of students according to X_+ serves as an estimate of their ordering according to θ (also see: Molenaar, 1982; van der Ark, 2005). Because the sum score is used in place of latent variable estimates on an interval scale (as in the case of parametric IRT), the MH model is described as an ordinal model for persons.

Monotonicity. In order to investigate monotonicity for an item of interest, it is necessary to examine the probability for a correct or positive response [$P(X_i=1)$] on the item across increasing levels of student achievement. This is accomplished by comparing the probability for success across increasing total scores (X_+). However, the interpretability of this comparison is challenged because the method used to calculate X_+ includes the item of interest. As a type of purification, Junker (1993) proposed the use of the *restscore* to check nonparametric model assumptions. A restscore is simply the total score (X_+) minus the score on the item of interest. To increase analytic power, students with adjacent restscores are combined into *restscore groups* that contain students with similar levels of achievement. Using these restscore groups, the assumption of monotonicity can be empirically checked. An item meets the monotonicity assumption of the MH model when the observed probability for a positive or correct response increases as student restscores increase. Methods for investigating monotonicity in empirical data, including statistical tests for the significance of violations, are described in detail in the presentation of the Monotone Homogeneity for Ratings model later in this chapter.

Ordinal Person Measurement. A set of items that meets the assumptions of the MH model is called a *monotonely homogeneous* item set. With the exception of ties, the expected order of students on the latent variable is the same across each item selected from a monotonely homogeneous set (Molenaar & Sijtsma, 2000). This "item-free" ordering of students allows inferences to be made about their ordering on the latent variable through the use of total scores. Adherence to the underlying MH model assumptions of unidimensionality, local independence, and monotonicity results in *ordinal person measurement*, or the stochastic ordering of person locations (SOL) on the latent variable (θ) by their observed total score when an instrument is made up of dichotomous items (Hemker, Sijtsma, Molenaar, & Junker, 1996; Huynh, 1994; van der Ark, 2012). The concept of SOL can be defined as follows: Two students *A* and *B* who can be ordered by their total scores (X_+) such that $0 \leq A < B \leq k$ can be ordered on θ by means of their total scores. Specifically, for a given fixed value: $\theta = c$, students A and B are ordered as follows under the MH model:

$$P(\theta > c \mid X_+ = A) \leq P(\theta > c \mid X_+ = B) \quad . \tag{26}$$

As pointed out by Molenaar & Sijtsma (2002), the above inequality implies that person total scores (X_+) can be used to order persons in terms of their θ locations; in other words, the average θ value within a group of students with a higher total score $X_+ = B$ is at least as high as the average θ value within a group of students who have a lower total score $X_+ = A$. Therefore, the mean θ in groups A and B is ordered as follows:

$$E(\theta \mid X_+ = A) \le E(\theta \mid X_+ = B) \quad . \tag{27}$$

for values of X_+ such that $0 \le A < B \le k$. Further, Sijtsma and Molenaar (2002, p. 22) demonstrated that the reverse relationship between θ and X+ is also implied by the MH model: θ estimates can be used to order students in terms of X_+.

Scalability. In addition to empirical checks for monotonicity, Mokken scaling techniques for the MH model also focus on examining the impact of error on the overall strength of a scale. Although Mokken scale analysis is based on Guttman's (1950) work on scaling techniques, Mokken (1971) recognized that a deterministic model does not accommodate the probabilistic nature of person responses to items, and that it is necessary to develop a method for explaining the impact of error on empirical observations. In his words:

> Perfect scales and perfect items rarely exist in practice. One has to face the fact that the ideal, as usual, can only be approximated. Scaling procedures … will be obstructed by the fact that once the order of the items has been ascertained, numerous "imperfect" patterns will occur, due to responses that are in error. (p. 42)

The impact of error is described in nonparametric IRT models through the use of scalability coefficients. Conceptually, scalability coefficients can be considered along the same lines as model-data fit statistics within a parametric approach. These coefficients have also been compared to indices of item discrimination (Sijtsma, Meijer, & van der Ark, 2011). Scalability coefficients that are used with Mokken scale analysis are based on Loevinger's (1948) *H* coefficient. Essentially, the *H* coefficient describes the frequency and magnitude of Guttman-type errors across a set of items. As pointed out by van Schuur (2011), the *H* coefficient can be interpreted in the same fashion as Goodman and Kruskal's (1979) lambda (λ) coefficient.

Loevinger (1948) described this coefficient as a method for identifying problematic items that result from three major types of errors: (1) ambiguity, or a lack of consensus among respondents' understanding of an item, (2) inclusion of items that are biased, and (3) inclusion of items that measure a construct other than the one of interest. Loevinger (1948) claimed that these three errors contribute to observed Guttman errors. Noting that Guttman's (1950) original method for scale analysis used the frequency of errors to classify a set of items as "scalable" or "not scalable," and, eventually as "scalable," "quasi-scalable," or "not scalable," Loevinger (1948) advocated for a more precise method for describing measurement quality. In her words:

> The concept of homogeneity has been developed as an alternative to the concept of reliability, and the degree of homogeneity of a test, like the concept of reliability, is intended to be stated numerically. [Guttman's classification of scalability] is comparable to a proposal that all tests be classified as "reliable," "quasi-reliable," or "unreliable." There are at least two objections. What is reliable (or homogeneous, or scalable) enough for some purposes is not good enough for others. But more essentially, no scientific purpose can be served by introducing discontinuities into our vocabulary which do not correspond to discontinuities in our data. Guttman has not offered any evidence that the lines

> between scales, quasi-scales, and non-scales are drawn to correspond to gaps in the data. (pp. 512-513)

In practice, the *H* coefficient is applied within the nonparametric IRT framework as a discrimination index used to determine the precision of person ordering on the latent variable by means of X_+. In order to increase this precision, values of *H* are used to identify items for potential removal from a scale. The scalability coefficients that are typically examined in Mokken-based analyses include indices of the scalability of item pairs (H_{ij}), individual items (H_i), and the overall set of items in a survey or test (H). These scalability coefficients describe the deviation of an observed data structure from a perfect Guttman (i.e., scalogram) pattern. In the case of dichotomous items, deviations from Guttman ordering are identified by determining the overall difficulty ordering of a set of items based on the proportion of students who succeed on an item. Then, the relative ordering of items within all possible item pairs is examined in order to check for discrepancies with the overall ordering. Specifically, checks for Guttman errors are conducted using pairs of items, and these checks involve comparing the proportion of students who succeeded on each item ($X = 1$) to the proportion of students who did not succeed on each item ($X = 0$). Because the difficulty order for the two items is known by the total proportion correct, Guttman errors are defined as instances of success on the harder item combined with failure on the easier item. After the frequency of Guttman errors is identified, the errors are weighted by the expected cell frequency that would occur given marginal independence. Finally, the ratio of observed errors to the frequency of expected errors is calculated. The item pair scalability coefficient (H_{ij}) is calculated as one minus this observed-to-expected error ratio, and it can be stated as:

$$H_{ij} = 1 - \frac{F_{i,j}}{E_{i,j}}, \tag{28}$$

where

H_{ij} = Scalability of the item pair consisting of Item *i* and Item *j*,

F_{ij} = Frequency of observed Guttman errors between Item *i* and Item *j*, and

E_{ij}= Expected frequency of Guttman errors between Item *i* and Item *j*, given marginal independence of the two items.

Sijtsma & Molenaar (2002) demonstrate a derivation of *H* coefficients as a ratio of the observed correlation between two items to the highest possible correlation, given the marginal distributions of the two items. For pairs of items, scalability is calculated using the covariance method as follows:

$$H_{ij} = \frac{\text{cov}(X_i, X_j)}{\text{cov}(X_i, X_j)^{\max}} \quad . \tag{29}$$

The numerator, which is the observed covariance between dichotomous Item *i* and dichotomous Item *j*, is calculated using the probabilities for each item:

$$Cov(X_i, X_j) = P_{ij} - P_iP_j \quad . \tag{30}$$

When $P_i < P_j$, and no Guttman errors are observed, the frequency of P_{ij} is equal to $P_{i.}$ Thus, the maximum covariance between dichotomous Items *i* and *j* given their marginals is:

$$Cov(X_i, X_j) = P_i - P_iP_j \quad . \tag{31}$$

The *H* coefficient can be stated as:

$$H_{ij} = \frac{Cov(X_i, X_j)}{Cov_{\max}(X_i, X_j)} = \frac{P_{ij} - P_iP_j}{P_i - P_iP_j} = 1 - \frac{P_i - P_{ij}}{P_i(1 - P_j)} \quad . \tag{32}$$

The scalability coefficient for a single item can be calculated in an analogous fashion. The item rest score ($R_{(i)}$) is the sum score across the set of *k* items besides the one of interest ($k - 1$). With the Guttman error cell for the item pair (i, j) defined as $X_i = 1$ and $X_j = 0$ when $P_i < P_j$, and $X_j = 1$ and $X_i = 0$ when $P_j < P_i$, the item scalability coefficient can be stated as:

$$H_i = \frac{Cov(X_i, X_{R(i)})}{Cov_{\max}(X_i, X_{R(i)})} = \frac{\underset{j \neq i}{\Sigma}(P_{ij} - P_i P_j)}{\underset{j>i}{\Sigma}(P_i - P_i P_j) + \underset{j<i}{\Sigma}(P_j - P_i P_j)}$$
$$= 1 - \frac{\underset{j \neq i}{\Sigma}(P_i - P_{ij})}{\underset{j>i}{\Sigma}(1 - P_i) + \underset{j<i}{\Sigma}(1 - P_j)}. \qquad (33)$$

Finally, the overall scalability coefficient for a set of items is stated as:

$$H = \frac{\sum_i Cov(X_i, R_{(i)})}{\sum_i Cov_{\max}(X_i, R_{(i)})} \quad . \qquad (34)$$

Appendix A illustrates the computation of rater pair scalability coefficients using the observed and expected error method and the method based on covariances.

Interpreting H. The *H* coefficient is a summary statistic that describes the degree to which a set of items approximate a perfect Guttman scalogram pattern, in which case this coefficient would have a value of $H = 1.00$, because there would be no expected Guttman errors. The lowerbound for the *H* coefficient is $H = 0.00$, and most values of *H* that are observed in real data have an approximate range of $0.30 \leq H \leq 0.60$ (Mokken, 1997; Sijtsma, Meijer, & van der Ark, 2011). It is common practice within Mokken Scale Analysis to apply rule-of-thumb critical values for the *H* coefficient in order to evaluate the quality of a scale (Mokken, 1971; Molenaar & Sijtsma, 2000). Typically, the following criteria are applied:

$H \geq .50$: Strong scale;

$.40 \leq H < .50$: Medium scale;

$.30 \leq H < .40$: Weak scale.

Molenaar and Sijtsma (1984) described the interpretation of the Mokken scalability coefficient (*H*) terms of the well-known alpha coefficient (Cronbach's α). They observed that both of these coefficients can be expressed in terms of the number of observed Guttman errors in a set of responses. Emphasizing the fact that these two coefficients describe unique properties of data, Molenaar and Sijtsma summarized previous comparisons of these two indices of internal consistency, and concluded that the value of the *H* coefficient is greater than α when there are very few items (≤ ~ 10) and large variance in item difficulties, or in the case of extremely low Guttman error frequencies. Except for cases in which all items have equal difficulty, such that the values of *H* and α are equivalent, the value of α will almost always be greater than that of the *H* coefficient.

Scalability and Automated Item Selection. One of the major uses of the scalability coefficient is to select sets of items that demonstrate adherence to the assumptions of Mokken scaling—i.e., Mokken scales. Mokken's (1971) original presentation of his scaling procedures includes a bottom-up method for selecting items that meet the assumptions of the MH model. This item selection procedure is based on the lowerbound for the scalability coefficient of $H \geq 0.30$ to identify sets of homogeneous, or scalable, items. As pointed out by Mokken (1997), the use of 0.30 as a lowerbound for H_i provides a method for establishing "long and useful scales" (p. 361). Further, Meijer, Sijtsma, and Smid (1990) demonstrated that the criterion of 0.30 for H_i

as an item selection technique yields discriminating items. Based on these observations, computer applications that implement Mokken scale analysis have been developed to include an Automated Item Selection Procedure (AISP) that identifies sets of scalable items using *H* coefficients. Essentially, Mokken's (1971) item selection procedure, along with its more-recent adaptations, is an exploratory procedure that involves three major steps. First, the pair of items (*i*, *j*) that have the highest item pair scalability coefficient (H_{ij}), and whose item pair scalability coefficient is significantly higher than zero, is identified. Then, a third item is selected that correlates positively with the items already selected and that has an individual item scalability coefficient (H_{f}) significantly larger than zero and greater than Mokken's (1971) recommended lowerbound of $H \geq 0.30$. The second step is repeated until all eligible items have been selected. As Mokken scale analysis software has been developed over time, updates and refined algorithms for automated item selection have been implemented (Straat, van der Ark, & Sijtsma, 2013; van der Ark, 2013). It is not yet clear how AISP applies to the context of rater-mediated assessment.

Double Monotonicity

The second nonparametric IRT model used in this study is Mokken's (1971) Double Monotonicity (DM) Model. This model is based on the same underlying assumptions as the MH model, with the additional assumption of nonintersecting item response functions, and, in the case of polytomous items, nonintersecting category response functions. For dichotomous items, adherence to the assumption of nonintersecting item response functions results in an *invariant item ordering.* Invariant item ordering (IIO) implies that that the ordering of items in terms of difficulty does not depend on which students are used for the comparison. Similarly, IIO implies that the ordering of students does not depend on the particular item by which the students are ordered. A set of items that satisfies these assumptions is called a *doubly monotone* set. With the exception of ties, the expected order of items in a doubly monotone set is identical for each subgroup of persons, and the expected ordering of persons on the latent variable is the same for each item. The IIO assumption states that, if two items can be ordered in terms of difficulty such that Item *i* is more difficult than Item *j* for a fixed location on the latent variable (θ), then the expected score for a student with any location on the latent variable can be ordered such that:

$$E_i\,(X \mid \theta) \leq \mathrm{E}_j\,(X \mid \theta) \quad , \qquad (35)$$

where $E_\mathrm{i}(X \mid \theta)$ and $E_j(X \mid \theta)$ represent the conditional expected score (X) on Item *i* and Item *j*, respectively, given a value on the latent variable (θ). The expected ordering in Equation (35) implies that the probability of succeeding on Item *i* is less than or equal to that associated with Item *j* across the range of the latent variable (θ):

$$P_i(X \mid \theta) \leq P_j(X \mid \theta) \quad . \qquad (36)$$

In nonparametric IRT analyses, restscores (R) are used in place of theta estimates (θ), and IIO implies that:

$$P_\mathrm{i}(X \mid R) \leq P_\mathrm{j}(X \mid R) \quad , \qquad (37)$$

where $P_i(X \mid R)$ and $P_j(X \mid \mathrm{R})$ represent the probability succeeding on Item *i* and Item *j*, respectively across all values of restscore *R*.

The DM model for dichotomous data has been described as a nonparametric or ordinal version of the Rasch model because of these invariant ordering properties. Meijer, Sijtsma, and Smid (1990) demonstrate the relationships among the MH, DM, and Rasch model in terms of restrictiveness for model-data fit, and point out several advantages of using a nonparametric model that is based on the assumption of nonintersecting IRFs:

> In many testing applications, it often suffices to know the order of persons on an attribute (e.g., in selection problems). Therefore, the Mokken model of monotone homogeneity seems to be an attractive model for two reasons. First, ordinal measurement of persons is guaranteed when the model applies to the data. Second, the model is not as restrictive with respect to empirical data as are the Mokken model of double monotonicity and the Rasch model. If, in addition an invariant ordering of items is required for all examinees (e.g., in intelligence testing), the model of double monotonicity may be appropriate. (p. 297)

Reliability for dichotomous Mokken models. Within the framework of Mokken scaling, internal consistency plays an important role in the evaluation of measurement quality. There are two major methods for examining internal consistency based on Mokken's (1971) *Theory and Procedure for Scale Analysis*: 1) scalability, and 2) an adaptation of the classical reliability coefficient ($\rho_{xx'}$) based on double monotonicity. The concept of scalability was described earlier, and the relationship between this coefficient and the well-known coefficient alpha is described in detail by Molenaar and Sijtsma (1984). Mokken (1971) presented his adaptation of the classical reliability coefficient using dichotomous items. However, this coefficient has also been formulated for use with polytomous items (discussed later in the chapter).

Mokken's (1971) method for examining the reliability of a total score (in his words: the reliability of a simple score, or "$\rho(\underline{s})$" is based on the classical reliability coefficient, which is an estimate of the proportion of observed score variance attributable to true score variance (Crocker & Algina, 1986). Classical reliability is estimated through the product-moment correlation between two independent replications of the same item: item-level reliability, or $\rho(x_i)$ or between two replications of a set of parallel items (reliability of a test score). Despite the fact that a nonparametric correlation coefficient, such as Kendall's (1938) rank correlation, may seem more intuitive than a parametric coefficient in the context of nonparametric IRT, several desirable properties of the parametric coefficient justify its use in the context of Mokken scaling. Sijtsma and Molenaar (1987) pointed out that the use of a nonparametric correlation coefficient is less informative as a description of the overall replicability of person ordering than its parametric counterpart, and that "for any smooth and well spread frequency distribution of observed scores—which is desirable in many measurement contexts—the conclusions from rank and product moment correlations are almost equivalent" (p. 81). Further, they noted that the fact that classical reliability coefficients are useful for providing information about the quality of measures obtained within a specific population, and that the coefficient provides a useful method for calculating the standard error of measurement.

Mokken (1971) discussed difficulties associated with estimating reliability that arise from the fact that it is not possible to achieve independent results from repeated administrations of test items, and that parallel items are based on assumptions that are difficult to realize in practice. Using π_i to represent the proportion of positive responses to a dichotomous item (i.e., item difficulty) and π_{ii} to represent the unobservable proportion of positive responses to two independent replications of the same item, Mokken (1971) summarized these challenges:

> We may stress here the well-known fact that $\rho(\underline{x}_i)$ and hence also $\rho(\underline{s})$ are not manifest parameters; they are functions of π_{ii} which is not a manifest parameter and cannot be estimated from the data in the same way as the π_i and π_{ij} (for $i \neq j$).
>
> We must know the form of $\pi(\theta,\delta)$, the value of δ_i, and the form and parameters of the population distribution… before we can determine π_{ii}. (p. 145)

Noting that item difficulty parameters (δ_i) are not directly estimated within the nonparametric IRT approach, and the fact that test-retest reliability requires assumptions that may not be achievable in operational settings, Mokken (1971) concluded, “the $\rho(\underline{s})$-coefficient of reliability is not estimable” (p. 145).

Reliability Estimation Procedures. As a solution to the problem of unobservable parameters for estimating $\rho(\underline{s})$, Mokken (1971) proposed two procedures for approximating π_{ii} that he described as “admittedly equally crude” as a method for estimating reliability based on parallel forms and test-retest methods (p. 146). Molenaar and Sijtsma (1984) proposed a third method that extended Mokken’s (1971) procedures. All three of these reliability estimation procedures are based on the assumption that IRFs do not intersect, i.e., that the DM model holds. The DM model implies that the following inequality holds when items are ordered by increasing proportion-correct values (i.e., in order of decreasing difficulty):

$$\pi_{i,i-1} < \pi_i < \pi_{i,i+1} \, , \tag{38}$$

where

> $\pi_{i,i-1}$ = the joint proportion of students scoring ‘1’ on dichotomous Item i and the item that immediately precedes it when items are ordered according to increasing difficulty, and

$\pi_{i,i+1}$ = the joint proportion of students scoring '1' on dichotomous Item *i* and the item that immediately succeeds it when items are ordered according to increasing difficulty.

Mokken (1971) proposed two methods for estimating π_{ii} that involve the use of the adjacent joint proportions $\pi_{i,i-1}$ and $\pi_{i,i+1}$. The first procedure involves identifying the smaller value between:

$$|\pi_i - \pi_{i-1}| \quad , \tag{39}$$

and

$$|\pi_i - \pi_{i+1}| \quad . \tag{40}$$

If (39) is smaller than (40), π_{ii} is estimated using $\pi_{i,i-1}$:

$$\pi_{ii} \approx \frac{\pi_i \pi_{i,i-1}}{\pi_{i,i-1}} \quad . \tag{41}$$

If (40) is smaller than (39), π_{ii} is estimated using $\pi_{i,i+1}$:

$$\pi_{ii} \approx \frac{\pi_i \pi_{i,i+1}}{\pi_{i,i+1}} \quad . \tag{42}$$

The second procedure uses interpolation between $\pi_{i,i-1}$ and $\pi_{i,i+1}$ to estimate π_{ii}:

$$\pi_{ii} \approx \pi_{i,i-1} + \frac{\pi_i - \pi_{i-1}}{\pi_{i+1} - \pi_{i-1}}(\pi_{i,i+1} - \pi_{i,i-1}) \quad . \tag{43}$$

Molenaar and Sijtsma (1984) describe a third method for estimating π_{ii}. The first step in this procedure is to replace π_{ii} with $(1-\pi_{ii})$ – the proportion of "zeros" on a dichotomous item. As given by Molenaar and Sijtsma (1984), substituting $(1 - \pi_{ii})$ into Equations (41) and (42) yields:

$$\pi_{ii} \approx \frac{\pi_{i,i-1}(1-\pi_i)}{(1-\pi_{i-1})} + \frac{\pi_i(\pi_i - \pi_{i-1})}{(1-\pi_{i-1})} \quad , \tag{44}$$

and

$$\pi_{ii} \approx \frac{\pi_{i,i+1}(1-\pi_i)}{(1-\pi_{i+1})} + \frac{\pi_i(\pi_i - \pi_{i+1})}{(1-\pi_{i+1})} \quad . \tag{45}$$

Then, the average value for π_{ii} is taken across (41), (42), (44), and (45). Using a simulation study, Molenaar and Sijtsma (1984) found that estimates of π_{ii} based on this average are less biased than estimates based on the first two procedures. Specifically, estimates of π_{ii} based on (42) and (45) were found to underestimate reliability (i.e., result in negative bias), and estimates based on (41) and (44) were found to overestimate reliability (i.e., result in positive bias). As a result, the average across the four estimates cancels out the positive and negative biases.

Once an estimate for π_{ii} has been obtained, these π_{ii} estimates can be used to estimate reliability for a test score as follows (Sijtsma & Molenaar, 1987):

$$\rho_{xx'} = 1 - \frac{\sum_i (\pi_i - \pi_{ii})}{\sigma^2(X)} \quad . \tag{46}$$

Sijtsma and Molenaar (1987) examined the three methods for estimating π_{ii} using simulation studies. They also explored challenges related to empirical situations in which several items have the same proportion correct (π_i), and when the distance between (39) and (40) are equivalent. These authors found that all three Mokken-based methods for estimating reliability had smaller bias than traditional methods for estimating reliability, including Cronbach's (1951) coefficient alpha (α) and Guttman's (1945) lambda coefficient (λ). When interpreting Mokken-based reliability coefficients, it is important to remember that, unlike classical estimates of reliability, Mokken-based reliability procedures assume that data fit the DM model. Therefore, it is necessary to assess fit to the DM model before interpreting these reliability coefficients.

Mokken Scaling for Polytomous Ratings

Molenaar (1982, 1997) extended Mokken's (1971) method for scale analysis for polytomous data, and proposed the polytomous MH model and the polytomous DM model. Similar to parametric polytomous IRT models, these nonparametric IRT models are based on a conceptualization of rating scale categories as a series of dichotomous "steps," such that $m - 1$ separate response functions are specified for a rating scale that has m unique categories. Within the framework of nonparametric IRT, Category Response Functions (CRFs) are defined as the cumulative probability for a rating within a category [$P(X \geq k \mid \theta)$]. The sum of the m CRFs equals the IRF:

$$E(X_i \mid \theta) = \sum_{m=1}^{m} P(X \geq k \mid \theta) \quad . \tag{47}$$

Nonparametric CRFs based on Mokken scaling are constrained by the same nondecreasing order restriction as the dichotomous IRFs (Equation 26).

Polytomous MHM. The polytomous MH model is based on the same underlying unidimensionality, conditional independence, and monotonicity assumptions as its dichotomous counterpart. For polytomous data, monotonicity can be stated as:

- *Monotonicity*: The probability that a person will receive a rating in Category *k* or higher increases as their location on the latent variable increases.

In the case of a rating scale with *m* categories, the monotonicity assumption places the following restriction on the cumulative probabilities for a score in category *k* ($X = k$) within groups of students *A* and *B* whose restscores (*R*) can be ordered such that $R_A < R_B$:

$$P(X \geq k \mid R_A) \leq P(X \geq k \mid R_B) \quad , \tag{48}$$

for all $m - 1$ categories.

For polytomous items, a weaker version of stochastic ordering on the latent variable (weak SOL) is implied by the MH model (van der Ark & Bergsma, 2010). Weak SOL implies that the total score (X_+) can be used to divide a sample into a group of persons with high locations on the latent variable and a group of persons with low locations on the latent variable, such that respondents with the highest and lowest θ values can be identified by dividing X_+ into two groups.

Polytomous DM model. The polytomous DM model is based on the same assumptions of unidimensionality, conditional independence, and monotonicity as the dichotomous DM model. For polytomous data, the DM model assumes nonintersecting CRFs:

- *Nonintersecting CRFs*: The order of cumulative category probabilities is consistent across all levels of student achievement.

It is important to recognize that there is a discrepancy between the concept of nonintersecting CRFs and nonintersecting items overall (nonintersecting IRFs). Specifically, nonintersecting CRFs, which are implied by the polytomous DM model (Molenaar, 1982, 1997), describe items at the level of rating scale categories, while IIO refers to nonintersecting IRFs. Recent research has highlighted a discrepancy between the polytomous DM model (Molenaar, 1982, 1997) and the property of IIO that characterizes the dichotomous DM model (Mokken, 1971). Specifically, Sijtsma and Hemker (1998) observed that, when CRFs are aggregated to IRFs (Equation 47), IIO is not observed in some cases, even though the CRFs do not intersect. In order to ensure IIO for polytomous data, Sijtsma and Hemker proposed the Strong Double Monotonicity (SDM) model, which is based on the assumption of nonintersecting CRFs *and* the assumption of nonintersecting IRFs. If CRFs are ordered such that IIO occurs, then the SDM model holds. Later, Ligtvoet and his colleagues proposed a method called Manifest Invariant Item Ordering (MIIO) that can be used to investigate nonintersecting IRFs for polytomous items based on the average rating for polytomous items (Ligtvoet, van der Ark, Bergsma, & Sijtsma, 2011; Ligtvoet, van der Ark, Janneke, te Marvelde, & Sijtsma, 2010).

Mokken reliability for polytomous scores. Molenaar and Sijtsma (1988) extended their three methods for estimating reliability (Molenaar & Sijtsma, 1984) to the case of polytomous scores, such as ratings. The reliability statistic presented by Molenaar and Sijtsma (1988) is referred to as the *MS statistic* in Mokken literature. Computation of the MS statistic is based on a view of polytomous items as a series of adjacent dichotomous item steps that are passed as a student receives a rating in a higher category. For Rater *i* scoring in adjacent rating scale categories *g* and *h*, π_{gi} and π_{hi} can be used to represent adjacent proportions in the DM model. Based on the idea of independent replications of the same rating procedure, $\pi_{gi,hi}$ represents the probability for a score of at least *g* and at least *h* from Rater *i* over two independent replications of the rating procedure; the value of $\pi_{gi,hi}$ cannot be directly estimated.

Methods used to estimate $\pi_{gi,hi}$ are parallel to the approximations of π_{ii} for dichotomous items. Proportions of scores in neighboring rating scale categories are used in a similar fashion to $\pi_{i,i-1}$ and $\pi_{i,i+1}$ for dichotomous items in order to estimate $\pi_{gi,hi}$. van der Ark (2010) summarizes methods for estimating the MS statistic based on adjacent joint cumulative properties. When rating scale category steps are arranged in order of decreasing difficulty, four proportions are used to estimate $\pi_{gi,hi}$, which is denoted $\mathrm{P}_{r,c}$ to reflect the joint cumulative proportion in the cell with Row r and Column c:

1. The lower neighboring joint cumulative proportion: $\mathrm{P}_{lo} = \mathrm{P}_{r+1,c}$
2. The right-hand neighboring joint cumulative proportion: $\mathrm{P}_{ri} = \mathrm{P}_{r,c+1}$
3. The upper neighboring joint cumulative proportion: $\mathrm{P}_{up} = \mathrm{P}_{r\text{-}1,c}$
4. The left-hand neighboring joint cumulative proportion: $\mathrm{P}_{le} = \mathrm{P}_{r,c\text{-}1}$

Then, using these proportions, eight estimates for $\pi_{gi,hi}$ are calculated:

$$P^{(1)}_{r,c} = P_{lo}\frac{\mathrm{P}_{\mathrm{r}}}{\mathrm{P}_{\mathrm{r+1}}} \quad ; \tag{49}$$

$$P^{(2)}_{r,c} = P_{ri}\frac{\mathrm{P}_{c}}{\mathrm{P}_{c+1}} \quad ; \tag{50}$$

$$P^{(3)}_{r,c} = P_{up}\frac{\mathrm{P}_{\mathrm{r}}}{\mathrm{P}_{\mathrm{r-1}}} \quad ; \tag{51}$$

$$P^{(4)}_{r,c} = P_{le}\frac{\mathrm{P}_{c}}{\mathrm{P}_{c-1}} \quad ; \tag{52}$$

$$P^{(5)}_{r,c} = P_{lo}\frac{1-\mathrm{P}_{\mathrm{r}}}{1-\mathrm{P}_{\mathrm{r+1}}} - P_{c}\frac{\mathrm{P}_{\mathrm{r+1}}-P_{r}}{1-\mathrm{P}_{\mathrm{r+1}}} \quad ; \tag{53}$$

$$P^{(6)}_{r,c} = P_{ri}\frac{1-\mathrm{P}_{c}}{1-\mathrm{P}_{c+1}} - P_{c}\frac{\mathrm{P}_{\mathrm{r}}-P_{r-1}}{1-\mathrm{P}_{\mathrm{r-1}}} \quad ; \tag{54}$$

$$P^{(7)}_{r,c} = P_{up}\frac{1-\mathrm{P}_{r}}{1-\mathrm{P}_{r-1}} - P_{c}\frac{\mathrm{P}_{\mathrm{r}}-P_{r-1}}{1-\mathrm{P}_{\mathrm{r-1}}} \quad ; \tag{55}$$

$$P^{(8)}_{r,c} = P_{up}\frac{1-\mathrm{P}_{c}}{1-\mathrm{P}_{c-1}} - P_{r}\frac{\mathrm{P}_{c}-P_{c-1}}{1-\mathrm{P}_{c-1}} \quad . \tag{56}$$

Then, the average of (49) – (56) is used as an estimate of the joint cumulative probability $\pi_{gi,hi}$. This value is used to compute the reliability coefficient for polytomous scores:

$$\rho_{XX'} = \sum_{i=1}^{k}\sum_{g=1}^{m}\sum_{j=1}^{k}\sum_{h=1}^{m}(\pi_{gi,hi} - \pi_{gi}\pi_{hi}) / \sigma^2(X) \quad , \qquad (57)$$

where *m* represents the number of categories, and *k* represents the number of items.

Across various presentations of the MS statistic for polytomous items (Molenaar & Sijtsma, 1984, 1988; Sijtsma & Molenaar, 1987), it is noted that methods for estimating reliability through the use of adjacent item difficulties $\pi_{gi}\pi_{hi}$ become complicated when items have identical proportions (π values), because the ordering of items in terms of difficulty is unknown. Molenaar and Sijtsma (1988) provided a brief overview of an alternative estimation method for $\pi_{gi}\pi_{hi}$ that accommodates these identical proportions; however, they did not provide a detailed description of the alternative estimates and their solution was not incorporated into the most recent version of the proprietary software for Mokken scaling (MSP5.0; Molenaar & Sijtsma, 2000). van der Ark (2010) observed that the MS statistic that is provided by MSP5.0 is incorrect in circumstances with identical item proportions, and provides a detailed overview of estimation methods suited for these situations. Essentially, this method involves identifying two or more adjacent joint cumulative proportions that have matching marginal cumulative proportions, and defining these identical proportions as a *set*. Sets are considered in place of the cells for estimates of joint cumulative proportions. The specific equations for various situations in which alternative estimates are needed are given in van der Ark (2010), and these provisional measures for calculating the MS statistic are incorporated into the *mokken* package for the *R* computer program (van der Ark, 2013; *R* Development Core Team, 2013).

Using Mokken Scaling to Examine the Quality of Ratings

Mokken's (1971) approach to nonparametric IRT for dichotomous data, and Molenaar's (1982, 1997) extension of these procedures to polytomous data have been widely applied to examine the measurement properties of affective measures across disciplines, including political opinions and participation (Mokken, 1971; Scarritt, 1996; van Schuur & Vis, 2000), public health (Sijtsma, Emons, Bouwmeester, Nyklicek, & Roorda, 2008), economics (Zinn, Henderson, Nystuen, & Drake, 1992), and psychiatric research (Bech, Hansen, & Kessing, 2006; Licht, Qvitzau, Allerup, & Bech, 2005; Watson, Deary, & Shipley, 2008). However, the utility of nonparametric IRT models for examining the quality of ratings assigned in an achievement test context has not been explored. The previously unexplored application of nonparametric IRT models to rater-mediated educational assessments is promising in light of the fact that parametric models may be unduly restrictive for rater-assigned scores, which may or may not adhere to the functional form requirements of the logistic ogive. Because they are less restrictive, nonparametric models that meet the requirements for invariant measurement may provide useful information that inform the use and interpretation of ratings in educational settings.

In this section, the polytomous versions of Mokken's (1971) MH and DM models (Molenaar, 1982, 1997; Sijtsma & Hemker, 1998) are extended for use as methods for examining the quality of ratings in the context of a large-scale rater-mediated educational assessment. The MH model is extended to the Monotone Homogeneity for Ratings (MH-R) model. Likewise, the SDM model is extended to the Double Monotonicity for Ratings (DM-R) model. Using these models, rating quality indices are adapted from the methods that are typically used to investigate measurement quality based on Mokken scaling for polytomous data. Table 13 summarizes rating quality indices based on the MH-R and DM-R models. As shown in the table, there are four major categories of rating quality indices based on Mokken scale analysis: A) rater scalability, B) rater monotonicity, C) rater double monotonicity, and D) invariant rater ordering. These rating quality indices are illustrated using analyses with the Georgia writing data.

Model III: Monotone Homogeneity for Ratings (MH-R) Model

This study presents the Monotone Homogeneity Model for Ratings (MH-R) Model, which is an extension of Mokken's (1971) MH model for use with polytomous rater-assigned scores. Methods that are typically applied within the nonparametric IRT framework to evaluate measurement quality based on the polytomous version of the MH model (Molenaar, 1982, 1997) are used here to examine the quality of rater-assigned scores. The next section describes two major rating quality indices based on the MH-R Model: A) scalability and B) monotonicity.

A. Scalability

The first indicator of rating quality based on the MH-R model is rater scalability. When the MH-R model is applied to polytomous ratings, scalability coefficients describe the degree to which a set of raters can be ordered to form a meaningful scale that describes differences among students in terms of a latent variable. Mokken (1971) proposed a minimum value of $H_i = 0.30$ to identify items that contribute to a meaningful ordering of persons in terms of a latent variable. Based on a view of raters as assessment opportunities, this critical value can be used to "flag" raters for further investigation. When the Mokken scalability coefficients are extended to polytomous ratings, deviations from a perfect Guttman ordering are identified by determining the relative difficulty ordering of rating scale categories across pairs of raters. Errors are defined as violations of this ordering that are observed in contingency tables for rater pairs. Then, the errors are weighted by their expected cell frequency that would occur given marginal independence. Finally, the ratio of observed errors to the expected errors is calculated. The rater pair scalability coefficient (H_{ij}) is calculated as one minus this observed-to-expected error ratio, in the same manner as presented in Equation 31. In a parallel fashion to item scalability coefficients, scalability coefficients for rater pairs (H_{ij}), individual raters (H_i), and a group of raters (*H*) can also be calculated using the covariance method (Sijtsma & Molenaar, 2002). Further, Mokken's (1971) criteria for classifying *H* coefficients provide a potentially useful index of the approximation to a Guttman scale within a set of polytomous rating data.

Results: Georgia writing data. Using the example dataset, the *H* coefficient for overall rater scalability was investigated for the 20 operational raters. Based on Mokken's (1971) critical values for the overall scalability coefficient, this group of raters appears to form a strong Mokken scale (*H* = 0.77; *SE* = 0.01). Values of individual rater scalability coefficients are presented in Table 14, Column A. Findings suggest differences in the relative frequency of Guttman errors across the group of raters. The lowest rater scalability coefficient is observed for Raters 7 and 21 (*H* = 0.74, *SE* = 0.02); nonetheless, this scalability coefficient suggests that the ratings assigned by these raters form a strong scale. The highest rater scalability coefficient is observed for Rater 9 (*H* = 0.82, *SE* = 0.02). The third step in the rater scalability analysis is an examination of the scalability of rater pairs. The H_{ij} coefficient is the normed covariance between ratings assigned by two raters, and positive values suggest adherence to the assumptions of the MH-R model by a pair of raters. For the 20 raters, there are 190 possible pairs. Results from the rater pair scalability analysis revealed that there were no negative rater pair scalability coefficients among the group of raters who scored the Georgia High School Writing Test.

B. Monotonicity

The second indicator of rating quality based on the MH-R model is monotonicity of ratings across the latent variable. In the case of polytomous ratings, the MH-R Model assumption of monotonicity states that the probability that a student will “pass” a rating scale category step (receive a rating in the higher of two adjacent rating scale categories) increases as their location on the latent variable increases. In order to determine whether an individual rater demonstrates the MH-R model assumption of monotonicity, it is necessary to examine the cumulative probability for each rating scale category on the item of interest [P($X \geq k$)] across increasing levels of student achievement. This is accomplished by examining the cumulative probability for ratings from a rater of interest across increasing restscores. A rater meets the monotonicity assumption of the MH-R Model when the observed cumulative probability for a rating in a higher category increases as student restscores increase.

Indicators of rater monotonicity include graphical displays and statistics that describe the degree to which a group of raters meets the MH-R model assumption of monotonicity. In this study, rater monotonicity is investigated using the *check.monotonicity* procedure in the *mokken* package (van der Ark, 2013), which examines monotonicity for individual raters using student restscores. In the context of a rater-mediated assessment, restscores are created using student total scores (X_+) across a group of raters minus the rating assigned by a rater of interest (X_i), such that $R = X_+ - X_i$. Then, this procedure combines students with adjacent restscores into restscore groups following the criteria for the minimum sample size within each group proposed by Molenaar and Sijtsma (2000, p. 67). Using restscore groups, the monotonicity assumption is investigated in two ways. First, monotonicity is examined in terms of average ratings within restscore groups. For each rater, the average rating within each restscore group is calculated. Increasing average ratings within increasing restscore groups provides evidence for monotonicity. Figure 17, Panel A is an example of a diagnostic plot that can be used to investigate monotonicity at the overall rater level. In this figure, student restscores are plotted along the *x*-axis, and average ratings on a four-point rating scale (0 = *low;* 3 = *high*) are plotted along the *y*-axis. This figure illustrates evidence of monotonicity for a single rater, because average ratings increase as restscores increase. Next, monotonicity is examined at the rating scale category level. Based on the conceptualization of polytomous ratings as a series of dichotomous steps, monotonicity is examined for the $m - 1$ meaningful category response functions by calculating the cumulative probability for a rating in a given category within each restscore group. If a rater demonstrates monotonicity, the cumulative probability for ratings in each category will increase as restscores increase. Figure 17, Panel B illustrates a diagnostic plot for rater monotonicity at the level of rating scale categories using a rating scale with four categories (0 = *low;* 3 = *high*). Student restscores are plotted along the *x*-axis, and the *y*-axis represents the probability for a rating in Category *k* or higher, given a restscore value [$P(X \geq k| R = r)$]. The highest line represents the probability that a student in a restscore group receives a rating in Category ‘1’ or higher [$P (X \geq 1)$]. Likewise, the second-highest line represents the probability for a rating of ‘2’ or higher [P ($X \geq 2$)], and the lowest line represents the probability for a rating of ‘3’ or higher [P ($X \geq 3$)].

Violations of monotonicity occur when the average rating in two adjacent restscore groups is disordered, such that students in the higher restscore group have a lower average rating than the students in the lower restscore group. The *mokken* package (van der Ark, 2013) tests the significance of violations of monotonicity at the overall rater level using values of *z*-test statistics. Specifically, a one-sided one-sample *z*-test is performed for the null hypothesis that the expected average ratings are equal between two adjacent restscore groups, against the alternative hypothesis that the expected average rating is lower in the group with a higher restscore, which would be a violation of monotonicity.

Results: Georgia writing data. In order to illustrate the second indicator of rating quality based on the MH-R model, monotonicity was examined for the 20 operational raters who scored the Georgia High School Writing Test. First, restscore groups specific to each rater were calculated for each of the 365 students. Because the highest rating from each rater is $X = 3$, the highest possible total score (X_+) for each student across the 20 raters is $X_+ = 60$. Thus, the highest possible restscore is $R = 57$, for students with the maximum score [$R = (X_+ – X_i) = (60 – 3)= 57$]. Using these restscore groups, rater monotonicity was examined at the overall rater level by comparing average ratings from a rater of interest across increasing restscore groups. Figure 18 illustrates rater monotonicity for the Georgia writing data at the overall rater level. Each plot describes monotonicity for a single operational rater. Restscore groups are plotted along the *x*-axis, and average ratings are plotted along the *y*-axis. Rater monotonicity is implied when average ratings increase as restscores increase. Next, monotonicity was examined for the 20 operational raters in terms of rating scale categories. Figure 19 illustrates the cumulative probability for ratings in each category using the three meaningful category response functions that correspond to the four-category rating scale. The three lines represent the three meaningful category response functions for the four-category rating scale. The lowest line is the cumulative probability for a rating in Category 3, the middle line is the cumulative probability for a rating in Category 2, and the highest line is the probability for a rating in Category 1. In general, the monotonicity plots in Figure 18 and Figure 19 suggest that this group of raters meet the MH-R model assumption for monotonicity.

In addition to graphical displays of monotonicity, the *mokken* package (van der Ark, 2013) can also be used to examine the statistical significance of violations of monotonicity. For each rating scale category *k*, the proportion of students in a restscore group who received a rating of X = *k* is compared to the proportion of students in each of the other restscore groups who received a rating in Category *k*. Appropriate comparisons between restscore groups are defined as comparisons between groups whose proportion of ratings in Category *k* is greater than zero (van der Ark, 2013). For each rater, these comparisons are performed for each rating scale category ($k = 1,\ldots,m$). Violations of monotonicity are identified during these comparisons when the probability that X = *k* is higher for the restscore group with a lower scale score. The significance of these violations is examined using *z* tests to compare proportions within the groups. Results from statistical tests for monotonicity for the Georgia writing data are summarized in

Table 14, Column B, which indicates that there are no significant violations of monotonicity for the 20 operational raters.

Model IV: Double Monotonicity for Ratings (DM-R) model

The next nonparametric IRT model used in this study is the Double Monotonicity for Ratings (DM-R) model, which is an adaptation of Sijtsma and Hemker's (1998) SDM model for use with polytomous rater-assigned scores. Methods that are typically applied within the nonparametric IRT framework to evaluate measurement quality based on the SDM model are presented here as methods to evaluate the quality of rater-assigned scores. The next section describes the remaining nonparametric rating quality indices in Table 13 that correspond to the DM-R model: C) double monotonicity, and D) invariant ordering.

In the context of a rater-mediated assessment, strong double monotonicity implies the following:

- *Nonintersecting category response functions*: Rating scale categories for individual raters have the same relative difficulty order across the range of the latent variable.

For example, if a rating in Category '*k*' from Rater *i* is more difficult than a rating in Category '*m*' from Rater *j*, then the probability for ratings in these two categories can be ordered such that:

$$P_i\,(X = k \mid R) < P_j\,(X = m \mid R)\ , \tag{58}$$

for all values *r* of rest score *R*. When adherence to the ordering assumption in Equation 58 is observed, rater severity ordering within categories can be interpreted in the same way for an entire group of students. In other words, non-intersecting cumulative category probabilities across raters imply that a set of rating scale categories has a meaningful order across the latent variable. This DM-R model assumption is based on the view of polytomous rating scales as a series of dichotomous "steps." When cumulative probabilities do not intersect across raters, these dichotomous steps have the same order across all levels of the raw score scale, such that the interpretation of rater severity does not depend on the student's location on the latent variable.

The DM-R model also implies:

- *Nonintersecting rater response functions:* The severity of individual raters is the same across students.

The assumption of nonintersecting RRFs implies that if the average ratings assigned by Rater *i* and Rater *j* are ordered for a student with restscore *R* such that $E\ (X_i) < E\ (X_j)$, then these two raters are ordered as follows under the DM-R model:

$$E(X_i \mid R) \leq E(X_j \mid R) \quad , \tag{59}$$

for all values *r* of rest score *R*. Indicators of rating quality for the DM-R model include graphical displays and statistics that describe the degree to which a group of raters meet the assumptions of strong double monotonicity.

C. Rater Double Monotonicity

The next indicator of rating quality based on Mokken scaling is rater double monotonicity, or nonintersecting CRFs. Adherence to rater double monotonicity can be investigated using several different procedures. In this chapter, two major methods for determining whether raters demonstrate double monotonicity are described: 1) rater restscore plots, and 2) proportion matrices.

Rater restscore plots. First, the assumption of nonintersecting CRFs can be checked in empirical data by comparing CRFs for pairs of raters across increasing restscore groups. This method is implemented in the *mokken* package for *R* (van der Ark, 2013) with the *check.restscore* function. Figure 20 illustrates graphical displays that can be used to investigate the assumption of double monotonicity for two raters. In both panels, student restscore groups are plotted along the *x*-axis, and the cumulative category probabilities for a four-category scale (0 = *low*; 3 = *high*) are plotted along the *y*-axis. Panel A demonstrates a violation of double monotonicity for Rater *i* (solid line) and Rater *j* (dashed line) related to rating scale categories '1' and '2.' Starting with category '1, ' the DM-R model assumption of non-intersecting cumulative category probabilities is violated because Rater *j* is more severe in this category than Rater *i* for the first restscore group, and the opposite is true for the second, third, and fourth restscore groups. A similar disordering is shown for the rating scale category '2.' In contrast, Panel B illustrates adherence to the DM-R model for Rater *i* and Rater *j*, whose relative severities within categories are consistent across the range of raw scores.

Results: Georgia writing data. Using the example dataset, the DM-R model was examined for each possible rater pair among the 20 operational raters who scored the Georgia High School Writing Test using graphical displays and significance tests to identify violations of double monotonicity. Figure 21 illustrates the graphical method to empirically check double monotonicity for two of the rater pairs involving Rater 14. Panel A illustrates adherence to the assumption of double monotonicity for Rater 14 and Rater 15, while Panel B illustrates violation to the DM-R model for Rater 14 and Rater 21.

In addition to graphical methods for examining adherence to the DM-R model assumption of double monotonicity, hypothesis tests can be calculated in order to determine whether violations of the DM-R model are statistically significant. The *mokken* package (van der Ark, 2013) does not provide significance tests for violations of nonintersection at the level of rating scale categories; instead, the significance of nonintersection is examined at the overall rater level. Table 14 summarizes results from the DM-R analyses for the Georgia writing data. As can be seen in the table, significant violations of double monotonicity were observed for six raters (Rater 5, 20, 4, 16, 17, and 21) via the restscore method.

Proportion Matrices. The third technique for investigating nonintersecting CRFs is known as the proportion matrix, or P (+, +)/ P (−,−) matrix, method (Mokken, 1971). In order to apply this technique, two matrices are created based on joint probabilities for "passing" rating scale category steps: the P (+, +) and P (−,−) matrices. First, the P (+, +) matrix represents the joint proportions of students receiving a rating in a particular category (i.e., a '1' on a rating scale category step) from two raters, P_{ij} (1,1). Second, the P (−,−) matrix describes the joint proportions of students *not* receiving a rating in a particular category (i.e., a '0' on a rating scale category step) from two raters, P_{ij} (0,0). In order to use the proportion matrix method, the rows and columns of P(+, +) and P(−,−) matrices are ordered by increasing probability values. Evidence for rater double monotonicity is apparent when it can be shown that the rows and columns of the P(+, +) matrix are non-decreasing, and that the rows and columns of the P(−,−) matrix are non-increasing.

Graphical displays within the *mokken* package (van der Ark, 2013) facilitate investigation of these matrices. The statistical significance of violations of non-decreasing P (+, +) and non-increasing P (–,–) matrices are calculated in order to identify raters whose violations of IRO warrant further attention. Figure 22 illustrates P (+,+) and P(–,–) matrices for each two raters that display the joint probabilities for observing ratings in a particular category between each rater and the remaining raters. In order to create the *P* (+, +) displays for an individual rater, the individual raters besides the one of interest are ordered on the *x*-axis from left to right in terms of decreasing severity. Then, the joint probability for observing a rating in a given category between the rater of interest and the rater on the *x*-axis is plotted. Evidence of rater double monotonicity is provided when the joint probability an observed rating increases as raters become less severe, which is indicated by an increasing line in the *P* (+,+) matrix plot. Likewise, the P (+, +)/ P (–,–) matrix plot suggests rater double monotonicity when the joint probability for an observed rating decreases as raters become more severe; evidence for rater double monotonicity is provided by a decreasing line in the P (–,–) matrix plot.

Results: Georgia writing data. Figure 23 and Figure 24 present P(+,+) and P(–,–) matrix plots for the Georgia writing data, respectively. In each plot of the P(+,+) matrices (Figure 23), the *x*-axis is the ordering of the rating scale category steps in terms of difficulty for the other 19 raters in the sample (20 raters minus the rater of interest = 19 raters x 3 rating scale categories = 57 steps shown along the *x* axis); the steps are ordered from most-difficult to most-easy. The P(–,–) plots (Figure 24) are similar, except the *x*-axis displays the rating scale category steps in order from easy to difficult. In both figures, the three plotted lines represent the joint probability for a rating in Category 2 (rather than Category 1; lower line), a rating in Category 3 (rather than Category 2; middle line), and a rating in category 4 (rather than 2; highest line). A violation of rater double monotonicity is implied when the lines in the P(+,+) matrix do not increase, and when the lines in the P(–,–) matrix do not decrease. Findings from statistical analyses of rater double monotonicity via the P(++)/P(–,–) method are summarized in Table 14. Significant violations of rater double monotonicity are observed most frequently for the ratings assigned by Rater 18, according to the proportion matrix procedure.

D. Invariant Ordering

The next indicator of rating quality based on the DM-R model is based on the SDM model (Sijtsma & Hemker, 1998) assumptions of invariant ordering for overall items and rating scale categories. As mentioned above, the mismatch between IIO from the dichotomous DM model and invariant item ordering for polytomous data has been addressed in recent research through the development of methods to investigate whether CRFs are ordered such that IIO is implied. Based on the interpretation of IIO at the overall item level (rather than within rating scale categories), Ligtvoet and his colleagues (2010, 2011) proposed a method called Manifest Invariant Item Ordering (MIIO) to check the assumption of IIO in polytomous data using average scores on items. As demonstrated by Ligtvoet et al., the MIIO method is used to investigate nonintersection through a combination of statistical and graphical techniques.

Extended to the context of a rater-mediated assessment, Ligtvoet et al.'s (2010, 2011) method will be referred to as *Manifest Invariant Rater Ordering* (MIRO) in this study. The MIRO procedure for examining nonintersection involves two major steps. First, the raters are ordered in terms of severity (i.e., difficulty) by their mean ratings across the entire group of students. Second, Equation 59 (nonintersecting RRFs) is evaluated for each pair of raters across restscore groups. Violations of IRO are apparent when rater severity ordering shifts across high and low restscore groups. If a violation is observed, a one-sided one-sample *t*-test is used to determine whether or not the reversal of rater severity is significant. Intersecting RRFs over values of restscores imply violations of IRO. Because the adaptation of Ligtvoet et al.'s (2010, 2011) method considers invariant rater ordering using average ratings, rather than focusing on individual rating scale categories, the MIRO method can be considered a more "crude," or general method for examining invariant ordering that describes overall raters (R. Ligtvoet, personal communication, July 18, 2013).

When the goal of a data analysis is to identify items that meet the assumption of MIIO, Ligtvoet et al. (2010, 2011) recommended that the MIIO method be used in a stepwise fashion and followed by the computation of a person scalability coefficient (H^{T}) as a measure of person fit. First, the MIIO method is applied to the entire set of items in order to identify items involved in violations of IIO. Then, these items should be removed, and the MIIO method repeated on the remaining items. This process continues until no violations of IIO are observed. When the final set of items is identified, the H^{T} coefficient is calculated, which is the same as the overall H coefficient, but calculated on the transposed data matrix. The H^{T} coefficient was first proposed by Sijtsma and Meijer (1992) as an indicator of overall model-data fit for the DM model. The range of this coefficient is $0 \leq H^{\mathrm{T}} \leq 1$, and higher values of H^{T} suggest adherence to the invariant person ordering property of the DM model—i.e., the order of items is invariant across a group of persons. Negative values of H^{T} indicate violations of the nonintersection assumption. In a parallel fashion to scalability coefficients for items, H^{T} describes the degree to which person ordering is free from Guttman errors; thus, Sijtsma and Meijer suggested that this coefficient be interpreted using the same critical values as the item scalability coefficients (H_{i}, H_{ij}, and H). However, as Ligtvoet et al. (2010, 2011) and Sijtsma et al. (2011) pointed out, H^{T} is difficult to interpret because it is sensitive to many different properties of a dataset. For example, the value of the H^{T} statistic is naturally higher when IRFs are farther apart; this information may or may not contribute to the selection of items that are appropriate for a given purpose. Sijtsma et al. summarized the challenges associated with the use of H^{T}, and concluded that this coefficient should not be interpreted as an indicator of measurement quality. As a result, this coefficient is not explored with the example dataset in this chapter.

In this study, graphical and statistical evidence is used to evaluate IRO. Specifically, pairwise plots are examined for pairs of raters in order to identify intersecting rater response functions. Figure 25 illustrates the graphical technique for examining MIRO with polytomous rating data. In Panel A, Rater *j* (dashed line) is more severe (lower expected ratings) than Rater *i* (solid line) for all restscore groups. On the other hand, the severity ordering for the rater pair in Panel B cannot be interpreted consistently across the raw score scale. Using the *mokken* package (van der Ark, 2013), hypothesis tests can be used to determine the significance of intersections. For example, if the overall average ratings from Rater *i* and Rater *j* can be ordered such that $\bar{X}_i < \bar{X}_j$, a violation of this ordering is observed for a particular restscore group *r* when this ordering is reversed, such that $(\bar{X}_i \mid R = r) > (\bar{X}_j \mid R = r)$. The significance of this violation can be examined by testing the null hypothesis that the conditional mean ratings for the two raters are equal, $(\bar{X}_i \mid R = r) = (\bar{X}_j \mid R = r)$ against the alternative hypothesis of the reversed severity ordering, which is a violation of MIRO.

Although it is possible to identify violations of MIRO through the use of statistical and graphical techniques, interpretation of these violations is not straightforward. For example, if two RRFs intersect, it is difficult to determine whether the violation of MIRO is due to one rater, due to the second rater, or due to the particular essay that is being scored. In the context of items (rather than raters), van Schuur (2003) explains that these violations are typically used to diagnose problems with items that can be used to inform survey or test development. He writes:

> The concept of a model violation thus revolves around a triple of objects consisting of one subject and two items. The number of model violations in a data set is defined as the number of transitivity relations among all such triples that are violated. As both subjects and items are involved, it is possible to attribute model violations to either subjects or items. It is usually not clear from the data or the theory as to which attribution is more plausible. But it is generally in the interest of researchers to keep their (representative) samples intact, and not to draw conclusions unless they can be supported by the whole sample and generalized to a wider population. Researchers may also be interested in identifying the most prototypical indicators of a specific concept, and judge that deleting items that are not homogeneous to the rest contributes to better measurement. Researchers therefore usually attribute violations to items rather than subjects. (p. 148)

There are techniques based on Mokken scale analysis for examining model violations that result from person misfit (Meijer, 1994; Meijer & van Krimpen-Stoop, 2001). This study focuses on model violations that are attributed to raters as an indicator of rating quality.

Results: Georgia writing data. Using the example dataset, MIRO was examined for each possible rater pair among the 20 operational raters who scored the Georgia High School Writing Test using graphical displays and significance tests to identify violations of MIRO. Figure 26 illustrates empirical checks for double monotonicity for two of the rater pairs involving Rater 8 using graphical displays. Panel A illustrates adherence to MIRO for Rater 8 and Rater 9, while Panel B illustrates violation to the DM-R model for Rater 8 and Rater 16. In addition to graphical displays to check double monotonicity, hypothesis tests can be calculated in order to determine whether violations of the DM-R model are statistically significant. Table 14, Column D summarizes results from the MIRO analyses for the Georgia writing data. These results suggest that there are violations of MIRO for several raters who scored the Georgia High School Writing Test. The most violations of IRO are observed for Rater 16, whose comparisons with the other raters in this sample resulted in three significant intersecting rater response functions. It is important to note that this finding that Rater 16 is most frequently involved in violations of MIRO does not necessarily imply that this rater assigns low-quality ratings. Because pairwise comparisons are used to investigate MIRO, it is difficult to identify which rater within a pair contributes to a model violation.

E. Reliability

The last indicator of rating quality based on Mokken scale analysis is reliability. In practice, applications of Mokken scaling include the use of the MS statistic (Molenaar

& Sijtsma, 1988). For this study, the polytomous version of the MS statistic (Sijtsma & Molenaar, 1987) is applied in order to investigate reliability for a set of raters. Before the MS statistic can be interpreted, it is important to verify that a set of polytomous data do not violate the nonintersecting response function assumptions of the DM model. The above results indicated some violations of this assumption for the example dataset. As a result, interpretation of the results from the reliability analysis of these data may be misleading, and these results are not reported.

Summary

This chapter presented and illustrated a set of indicators of rating quality that can be applied in the context of rater-mediated educational assessments. This set of indicators is based on the indicators of measurement quality that are typically used in applications of Mokken scaling. Specifically, Mokken's (1971) MH and DM models, along with the polytomous versions of these models (Molenaar, 1982, 1997; Sijtsma & Hemker, 1992) were used to investigate rating quality in terms of scalability, monotone homogeneity, double monotonicity, invariant ordering, and reliability. As was demonstrated in this chapter, the interpretation of these Mokken-based indices is slightly different when they are used as indices of measurement quality with rater-assigned scores. Overall, the illustrative data analyses suggest that Mokken scale analysis provides useful information that can augment parametric methods for examining the requirements of invariant measurement, such as those provided by the Rasch model (Rasch, 1960/1980).

Chapter Three and Chapter Four extended the traditional rating quality indices described in Chapter Two to a set of rating quality indices based on two approaches to IRT: Rasch Measurement Theory (Rasch, 1960/1980) and Mokken scaling (Mokken, 1971). Next, Chapter Five is an empirical application of these techniques that also considers the degree to which indices of measurement quality based on a parametric and nonparametric approach lead to comparable conclusions. Then, Chapter Six presents a summary of the findings from these two empirical applications and draws conclusions for the entire dissertation based on the research questions for this study.

Chapter Five: Examining Rating Scales using Rasch and Mokken Models for Rater-Mediated Assessments

Chapter Five presents an empirical application of Rasch measurement theory (Rasch, 1960/1980) and Mokken scale analysis (Mokken, 1971) to data from a large-scale rater-mediated writing assessment. This chapter presents and illustrates a set of parametric and nonparametric indicators of measurement quality that complement the rating quality indices that were described in Chapter Three and Chapter Four. Specifically, the focus of this application is on the use of techniques from Rasch measurement theory and Mokken scaling to examine the quality of rating scales in a rater-mediated assessment. As a result, the focus of this chapter is slightly different from the emphasis on individual raters in Chapters Three and Four. In light of the view of rater-mediated assessments as ecological contexts in which a rater acts as a type of "lens" through which a performance is interpreted, it is also important to empirically investigate the structure of rating scales in terms of how individual rating categories are applied by the raters. The techniques presented in this chapter are adapted from a well-known set of guidelines for evaluating the effectiveness of rating scales presented by Linacre (1999, 2004). In addition to empirical illustrations of these techniques, the degree to which parametric and nonparametric indices of adherence to the guidelines lead to comparable conclusions about the quality of rating scales is considered.

Introduction

In general, the use of rater-mediated assessments reflects a view that a rater's judgment of a response provides information beyond what could be provided by a more objective measure, such as a set of multiple-choice items. However, the interpretation and use of rater-assigned scores for informing educational decisions depends on the degree to which rater-mediated measurement systems demonstrate useful psychometric properties. Operational methods for evaluating the quality of rater-mediated assessments include indices of rater agreement, errors, and accuracy (Murphy and Cleveland, 1991; Johnson, Penny, and Gordon, 2009). Further, methods for evaluating the quality of ratings have been developed based on true score and latent trait models that include indices of a variety of rater effects (e.g., Brennan, 1996, 2000; Engelhard, 2013; Wolfe, 2009). Although indicators of rater effects are useful for informing the interpretation and use of rater-assigned scores, it is also important to consider the quality of ratings in terms of the structure of the rating scale on which scores are assigned. Hamp-Lyons (2011) pointed out that rating scales for performance assessments serve two major roles. First, they are tools that can be used to compare a student's response to a set of performance criteria; Second, they serve as operational definitions of a construct, such as writing ability. In the context of writing assessment, she observed: "Often we think of rating scales as tools, and perhaps they are. But they are also realizations of theoretical constructs, of beliefs about what writing is and what matters about writing" (p. 3). Human raters combine these two purposes for rating scales when they use a set of performance criteria to describe a student in terms of a theoretical construct, such as writing achievement. In order to trust ratings as useful descriptions of student performance, it is necessary to systematically examine the degree to which a rating scale is functioning as intended for a group of students. Further, concern with the validity of rater-assigned scores necessitates investigation of these quality control indices within student subgroups, such as gender, language, and race/ethnicity groups. This chapter demonstrates methods that can be used to examine rating scale functioning within and across student subgroups based on parametric and nonparametric item response theory models for rater-mediated assessments.

Purpose

The major purpose of Chapter Five is to develop and explore diagnostic indicators for rating scales based on Mokken (1971) scaling. Specifically, indicators of rating scale effectiveness are explored based on the Monotone Homogeneity for Ratings (MH-R) model and the Double Monotonicity for Ratings (DM-R) model, both of which were presented in Chapter Four. In order to provide a frame of reference for considering the utility of nonparametric IRT indices, indicators from Mokken scaling are compared to indices from the Partial Credit (PC) model (Masters, 1982; Wright and Masters, 1982) that was presented in Chapter Three. The correspondence between the two approaches is examined in terms of conclusions about rating scale effectiveness. Further, the degree to which rating scale quality is invariant across gender and race/ethnicity subgroups is examined within each approach.

Research Questions

In this study, parametric and nonparametric IRT models are used to examine rating scale effectiveness in general, and within student subgroups. The first two research questions focus on indicators from each modeling approach separately:

1. What does Rasch measurement theory reveal about the quality of a rating scale for a large-scale rater-mediated writing assessment within and across student subgroups?
2. What does Mokken scale analysis reveal about the quality of a rating scale for a large-scale rater-mediated writing assessment within and across student subgroups?

The last research question focuses on the correspondence between indices of rating scale effectiveness between the parametric and nonparametric approaches:

3. Do indices of rating scale category effectiveness based on Rasch measurement theory and Mokken scale analysis provide comparable information about the overall quality of a rating scale?

Procedures

The research questions for this study are examined within the context of a large-scale rater-mediated writing assessment. In this section, the instrument and participants from whom empirical data were collected are described. Then, the methods used to perform the parametric and nonparametric analyses are presented.

Instrument

The instrument used in this study is the Alaska High School Graduation Qualifying (HSGQ) exam, which is first administered in Grade 10. The HSGQ Exam includes a combination of multiple-choice, short construct-response (SCR) and extended constructed-response (ECR) items in three subject areas: mathematics, reading, and writing. This chapter focuses on the ECR items in the writing section of the HSGQ Exam. Specifically, four ECR items that were included in Spring 2013 administration for tenth-grade students are used for the analyses. Three of the ECR items (ECR1, ECR3, and ECR4) are scored in four categories (1 = *low*; 4 = *high*), and one ECR item (ECR2) is scored in six categories (1 = *low*; 6 = *high*). For the purposes of the analyses in this chapter, these scales were recoded to (0 = *low*; 3 = *high*) for the four-category items, and to (0 = *low*; 5 = *high*) for the six-category item. Students received ratings from two raters on each ECR item. Although additional ratings were assigned in the case of rater disagreement as part of a score resolution technique during operational scoring, this study focuses on the two initial ratings assigned to each student for the four ECR items. IRB information for use of this data is provided in Appendix B.

Participants

The data used in this study are from a sample of students and raters who participated in the Spring 2013 administration of the writing portion of the tenth-grade HSGQ Exam in Alaska. Participants include 8,620 tenth-grade students whose responses to the ECR questions were scored by a group of 64 raters. Each rater scored at least 370 responses, and the raters formed a connected rating design (Eckes, 2009; Engelhard, 1997). In order to examine rating scale effectiveness in terms of student subgroups, data analysis procedures were applied separately for the students in the following subgroups: female (N = 4,218), male (N = 4,402), Alaskan native (N = 1,729), and White (N = 4,502).

Data Analysis

This study presents and illustrates a set of indicators of rating scale category effectiveness based on Rasch measurement theory (Rasch, 1960/1980) and Mokken scale analysis (Mokken, 1971). Further, the study focuses on the comparability of rating scales for the ECR items on the HSGQ Exam across students in the male and female gender subgroups and across students in the Alaskan native and White race/ethnicity subgroups. These subgroups were selected based on previous research on large-scale writing assessments that has identified persistent differences in student achievement based on gender and race/ethnicity (Engelhard, Wind, Kobrin, and Chajewski, in press).

The data analysis procedures for this study involved several steps. First, the Facets computer program (Linacre, 2010) was used to estimate the PC Rasch model for the overall student sample and within the gender and race/ethnicity subgroups. Likewise, the *mokken* package for the *R* statistical software program (van der Ark, 2013; *R* Development Core Team, 2013) is used to estimate the MH-R and DM-R models for the overall student sample and for the gender and race/ethnicity subgroups. Then, a set of guidelines for rating scale effectiveness (discussed below) was examined for the overall student sample and within the gender and race/ethnicity subgroups.

Guidelines for Rating Scales

Linacre (1999, 2004) presented a set of guidelines for evaluating the quality of rating scales. For this study, a set of rating scale guidelines adapted from Linacre are used to explore the structure of the rating scales for the ECR items. Linacre presented these criteria within the framework of Rasch measurement theory (Rasch, 1960/1980). However, he noted: "though these guidelines are presented within the context of Rasch analysis, they reflect aspects of rating scale functioning which impact all methods of analysis" (p. 85). In order to compare indices of rating scale functioning across the Rasch (parametric) and Mokken (nonparametric) models, Linacre's criteria are summarized into three major guidelines within which results from the Rasch- and Mokken-based models can be grouped in order to examine rating scale category effectiveness: 1) directional orientation with the latent variable, 2) category precision, and 3) model-data fit.

Table 15 summarizes these guidelines and presents indices of adherence based on the Rasch PC model (Masters, 1982), the MH-R model, and the DM-R model. The entries in this table constitute the methods used to evaluate the structure of the HSGQ

Exam for this study. In the next section, these guidelines are described in general, and they are illustrated as they apply to Rasch measurement theory and Mokken scale analysis. The visual displays and illustrative statistics related to the PC model were created using the Facets computer program (Linacre, 2010), and the displays and statistics related to the MH-R and DM-R models were created using the *mokken* package for *R* (van der Ark, 2013). Data analyses for this study include an examination of these three guidelines for the overall sample of students, and within the gender and race/ethnicity subgroups.

1. Directional Orientation with the Latent Variable

The first major guideline for rating scale categories is related to the orientation of sequential rating scale categories in the same direction as the latent variable. Adherence to this guideline suggests that increasing amounts of a latent variable (θ) correspond to increasing categories on a rating scale. In the context of a rater-mediated writing assessment, adherence to this guideline suggests that raters are interpreting the rating scale categories in the manner implied by the ordered categories. Table 15 indicates that there are several methods to examine this guideline based on Rasch measurement theory and Mokken scaling. The Rasch-based indicator is: A) expected score ogive has a positive slope. The indicators based on Mokken scale analysis include: B) average ratings increase monotonically across restscores, and C) category response functions increase monotonically across restscores.

Rasch-based Evidence for Orientation with the Latent Variable

The PC model can be used to examine empirical functioning of a rating scale for evidence of directional orientation with the latent variable. Specifically, expected ratings based on the PC model can be examined for evidence of adherence to this guideline.

A. Expected score ogive has a positive slope. Using the PC model (Equation 11; see Chapter Three), expected ratings can be calculated that correspond to student locations on the latent variable. When rating scale categories are oriented in the same direction as the latent variable, expected ratings increase along with student locations on the latent variable. Panel A in Figure 27 includes expected score ogives that illustrate this property for an item scored using a four-category rating scale (0 = *low;* 3 = *high*) under the PC model. Plotted along the *x*-axis are student measures relative to item difficulty calibrations ($\theta - \delta$). The *y*-axis shows expected ratings, which range from 0 to 3. As can be seen in the first plot, there is a positive relationship between values on the *x*-axis and values on the *y*-axis, which suggests adherence to the orientation guideline. In contrast, Panel B in Figure 27 displays evidence of minor violations of this guideline. Specifically, “dips” in the function that relates the *x*- and *y*-axes suggest a non-monotonic relationship between rating scale categories and locations on the latent variable.

Mokken-based Evidence for Orientation with the Latent Variable

Mokken scale analysis can also be used to examine the degree to which a set of ratings is oriented in the same direction as the latent variable. The two nonparametric IRT models used in this study are based on the assumption of monotonicity in the latent variable. Checks for adherence to the underlying monotonicity assumption for both the MH-R model and the DM-R model can be used to evaluate a set of ratings in terms of the directionality guideline. Because Mokken scaling does not impose enough restrictions on the functional form of the item response function to compute logit-scale estimates of person locations on the latent variable (θ), student restscores are used to investigate monotonicity. First, restscores (total score minus the score on an item of interest; see Chapter Four) specific to each item are calculated by subtracting student scores on the item of interest from their total score across items. Then, students with adjacent restscores are combined into groups in order to increase statistical power. Using these restscore groups, the monotonicity procedure in the *mokken* package (van der Ark, 2013) produces graphical and statistical evidence of monotonicity for each item overall (B), and for the rating scale categories within each item (C).

B. Average ratings increase monotonically across rest scores. The first indicator of directional orientation with the latent variable based on Mokken scaling is a monotonic increasing relationship between average ratings and student achievement on an item of interest. This indicator is demonstrated in Panel B of Figure 27 for a rating scale item with four ordered categories. Average ratings (*y*-axis) for students are plotted along increasing restscores (*x*-axis). The first plot in Panel B illustrates increasing average ratings for students with increasing restscores. In contrast, the second plot in Panel B illustrates a non-monotonic relationship between student restscores and average ratings. Values of test statistics are examined in order to determine whether violations of monotonicity are significant at the overall item level. Specifically, a one-sided one-sample *t*-test is performed for the null hypothesis that the expected average ratings are equal between two adjacent restscore groups (the boundary of permissible means under the MH model), against the alternative hypothesis that the expected average rating is lower in the group with a higher restscore, which would be a violation of monotonicity.

C. Category response functions increase monotonically across rest scores. The second Mokken-based indicator of directional orientation with the latent variable is related to the rating scale categories within each item. This indicator is demonstrated in Panel C of Figure 27 for a rating scale item with four ordered categories (0 = *low;* 3 = *high*). Based on the conceptualization of polytomous ratings as a series of dichotomous steps, these figures display the cumulative probability for a rating in a given category within each restscore group. The highest line represents the probability that a student in a restscore group receives a rating of '1' or higher, the middle line represents the probability for a rating of '2' or higher, and the lowest line represents the probability for a rating of '3.' Illustrated in the first display of Panel C, evidence for directional orientation with the latent variable is seen when these cumulative probabilities increase over increasing restscore groups. On the other hand, decreasing cumulative probabilities for a rating category suggest violations of this guideline; the second display in Panel C demonstrates a violation of this guideline.

2. Category Precision

The next guideline for evaluating rating scale functioning is related to the precision with which rating scale categories distinguish among students in terms of the latent variable. In order for rating scale categories to be interpretable, it is necessary that they reflect meaningful differences between students in terms of the latent variable. For example, in the case of a four-point rating scale (0 = *low*; 3 = *high*), this guideline requires that there is a meaningful difference between logit scale locations for students who receive a rating of ‘3’ and students who receive a rating of ‘2.’ Figure 28 summarizes methods based on Rasch measurement theory and Mokken scaling that can be used to evaluate this guideline; each entry in Figure 28 is elaborated below.

Rasch-based evidence for Category Precision

Within the framework of parametric IRT, the precision with which rating scale categories distinguish among students in terms of the latent variable is influenced by the relationship between student locations on the logit scale that represents the latent variable (θ) and the probability for scores in rating scale categories [P($X = k$)]. Indices of category precision based on the PC model include: A) the frequency distribution of students across rating scale categories, B) the location of category coefficient parameters on the logit scale, C) conditional category probabilities, and statistical information for overall items (D) and rating scale categories (E).

A. Normal or uniform distribution of students across rating scale categories. Linacre (2004) pointed out that the percent of observations within each rating scale category has important implications for the estimates of category coefficients (δ_{ij}). First, frequency distributions of students within rating scale categories can be inspected. The first plot in Figure 28, Panel A presents an example of a somewhat uniform rating distribution that satisfies this guideline; in contrast, the second plot is an example of a left-skewed rating distribution that violates the guideline. Because the estimation of category coefficient locations for parametric models depends on frequencies of observations within rating scale categories, Linacre (2004) has suggested as a rule of thumb that categories with fewer than ten observations limit the precision and stability of these estimates. Unobserved categories present significant challenges to the interpretation of rating scales. Categories with no observations must be distinguished as either structural or incidental zeroes. A structural zero occurs when category requirements are impossible to fulfill, and an incidental zero occurs when an unobserved category is the consequence of a particular sample. Linacre (2004) describes strategies for addressing issues related to unobserved categories.

B. Absolute value of the difference between each category coefficient (δ_{ij}) is between approximately 1.40 and 5.00 logits. This requirement is related to the distance between rating scale category coefficients on the logit scale. In order to describe meaningful differences among students in terms of the latent variable, it is necessary that each rating scale category describes a unique range of values on the latent variable. As a rule of thumb, Linacre (1999, 2004) proposed a minimum difference of about |1.40| logits between categories for rating scales with three categories, and a minimum difference of about |1.00| logit between categories for rating scales with five categories. Differences smaller than these minimum values suggest that there may not be a meaningful difference between rating scale categories in terms of the latent variable. On the other hand, Linacre (1999, 2004) proposed a maximum difference of about |5.00| logits between rating scale category coefficients. Differences larger than about five logits suggest that a rating scale category may mask meaningful differences among students in terms of the latent variable. Figure 28, Panel B illustrates two four-category rating scale items whose category coefficients meet and violate this guideline.

C. Multimodal category response functions. Figure 28, Panel C illustrates category probability functions from the PC model, which are the next indicator of category precision based on the PC model. Category probability functions are a visual representation of the probabilistic relationship between category difficulty and student location on the latent variable. Each curve represents an individual rating scale category, and the curves always appear in ascending order so that the curve representing the lowest category is farthest to the left and the curve for the highest category is farthest to the right. When rating scale categories effectively distinguish among students in terms of the latent variable, these functions appear as a "range of hills" – that is, each curve has a distinct peak (Linacre, 1999, 2004). Essentially, this guideline requires that each category describes a distinct range on the latent variable where it is the most likely. Graphical displays are useful for examining this guideline in order to gain a sense of the distinctiveness of each rating scale category within a particular dataset. A set of multimodal category response functions is illustrated in the first plot in Figure 28, Panel C. In contrast, the second plot in Figure 28, Panel C demonstrates a set of category response functions for which modality is not observed for each of the categories. Although the formulation of the category probability curves is such that they always appear in ascending order from left to right, the crossover points between adjacent curves can be disordered if the scale is not functioning as intended. As can be seen in the second plot in Figure 28, Panel C, rating scale categories that are never the most probable at any point along the *x*-axis (i.e., are non-modal) do not describe a unique range of the latent variable.

D. Conditional category probability curves are distinct and evenly spaced along the latent variable. The next indicator of rating scale category precision based on the PC model is related to the location of conditional category probability curves along the logit scale. Illustrative conditional category probability curves are given in Panel D of

Figure 28. These curves are logistic ogives that describe the conditional probability for "passing" adjacent rating scale category steps. Each curve represents two categories, such that the curve farthest to the left models the probability for a rating of the lowest and next-lowest categories across the range of the latent variable. In the Facets program (Linacre, 2010) output, a dashed horizontal line intersects each curve at the 0.50 probability point to indicate the location of the Rasch-Andrich threshold for each pair of categories: This is the point on the latent variable at which a category is most probable. Similar to category probability curves, evenly spaced and distinct conditional probability curves along the logit scale suggest that each rating scale category describes a unique range of locations along the latent variable.

E. Smooth item information function. Figure 28, Panel E displays information functions for two rating scale items. Item information functions provide diagnostic information about rating scale items because they display the amount of model-based (Fisher) statistical information provided by an item at different locations on the latent variable (Fisher, 1958). Item information is related to the match between person location and item difficulty, and well-targeted items provide more information than items that are far from person locations. Item information is directly related to the precision of measurement, such that measures with small standard errors contribute more information than measures with large standard errors. Whereas item information functions for dichotomous items are maximized at their logit-scale locations (δ_i), the distribution of information for polytomous items depends on the model that is specified. For the PC model, which is used in this chapter, item information is influenced by the distance between category coefficients and whether or not category coefficients are disordered (de

Ayala, 2009). The plot of item information for rating scale items can be used to identify locations along the latent variable at which the information is most useful for providing statistical information, identified by high values on the *y*-axis. The desired shape of item information functions for rating scale items varies depending on the purposes and consequences of a particular assessment. Generally, item information functions for achievement tests are inspected for "valleys," or substantial reductions in information across a specific range of person locations, that may signal a reduction in the precision with which students are described in terms of the latent variable. For rating scale items, reductions in information frequently occur when there are large distances between category coefficient locations. The first display in Panel E of Figure 28 illustrates a rating scale item that provides high values of information across a wide range of the latent variable. In contrast, the second display in Panel E illustrates a reduction in information for students who are located between about −3 and 0 logits.

F. Smooth category information functions. Similar to item information, category information functions also provide diagnostic information about the precision of rating scale categories. In Figure 28, Panel F, the amount of information provided across student locations on the latent variable is plotted separately for each rating scale category. When a rating scale is functioning as intended, lower categories will provide more information for students with low measures on the latent variable than for students with high measures. Reductions in category information are interpreted the same way as for overall items.

Mokken-based Evidence for Category Precision

Mokken scale analysis can also be used to examine the precision of rating scale categories. Similar to indices of category precision based on the parametric PC model, estimates of category precision based on nonparametric IRT models focus on the degree to which rating scale categories distinguish among students in terms of the latent variable. However, because nonparametric models do not provide interval-level estimates of student locations on the latent variable, estimates of category precision are based on cumulative probabilities for ratings using information from the raw score scale.

G. Category response functions do not overlap *within* items. Panel G in Figure 28 displays category response functions based on Mokken's (1971) nonparametric IRT models. This display was examined earlier as an indicator of category monotonicity. However, cumulative category probabilities based on nonparametric IRT can also be used to determine whether rating scale categories within an item distinguish among students in terms of the latent variable. When rating scale categories describe meaningful differences among students in terms of the latent variable, the probability for a rating in each category will be distinct. Graphical displays of cumulative category probabilities provide evidence for adherence to this guideline when the line that represents a rating scale category does not overlap with the line that represents another category. The first display in Panel G of Figure 28 demonstrates adherence to this guideline for a four-category rating scale item (0 = *low*; 3 = *high*). On the other hand, if rating scale categories do not indicate meaningful differences in terms of the latent variable, the cumulative probabilities will have similar values. The second display in Figure 28 Panel G illustrates overlapping cumulative probabilities for the first two categories for an item with a four-point rating scale. These overlapping cumulative probabilities suggest that the first two rating scale categories provide redundant information.

H. Category response functions do not overlap *across* items. Panel H of Figure 28 extends the previous indicator of category precision to multiple items. When cumulative category probabilities are considered for multiple items, non-overlapping cumulative probabilities *across* items suggest that a set of rating scale items function together to describe differences among students in terms of the latent variable. For example, the first display in Panel H illustrates two rating scale items whose categories provide distinct information. Specifically, the cumulative category probabilities for the solid-line item are distinct from those for the dashed-line item, which suggests that the items provide unique information. On the other hand, the second display indicates that the two rating scale items provide redundant information, because the cumulative probabilities for the rating scale categories are about the same for both items. It is interesting to note that although the situation in the second display does not violate the requirements of the DM-R model because the category probabilities do not intersect, these two items do not function together to provide distinct information about students that may be important in a rater-mediated assessment.

3. Model-data fit

The last major guideline that can be used to examine rating scale effectiveness is related to model-data fit. As given in Table 15, this guideline requires that rating scale categories meet the expectations of models with useful measurement properties, such as invariant measurement or invariant person and item ordering. Indicators of model-data fit for rating scale categories can be used to determine whether individual categories are functioning as intended by a measurement model. Figure 29 illustrates two indicators of model-data fit for the PC model and three indicators of model-data fit based on Mokken scale analysis that can be used to evaluate adherence to this guideline in empirical rating data.

Rasch-based Evidence of Model-Data Fit

Two indicators of model-data fit to the PC model are particularly useful for providing diagnostic information about the effectiveness of a rating scale. First, observed ratings for students at different locations on the logit scale can be compared to their expected ratings based on the PC model using a visual display. Next, model-data fit statistics that summarize residuals, or differences between model expectations and empirical observations can be used to identify individual rating categories for which rating patterns do not match the values predicted by the PC model.

A. Close match between observed and expected score ogives. First, the PC model can be used to examine the correspondence between empirical and expected ratings. When the PC model is applied to rating data, expected ratings are calculated for a range of values on the logit scale. Using these expected ratings, it is possible to create a visual display of expected and observed score ogives as a graphical method for examining model-data fit. Figure 29, Panel A includes a Rasch model expected score ogive that includes empirical observations. The Xs identify the observed average rating on the *y*-axis for an interval of student measures on the latent variable (*x*-axis). Confidence bands are drawn around the curve that represent upper and lower bounds of a 95% confidence interval. Observations that fall outside these bands indicate misfit, or unexplained variance. A close match between the expected and empirical ratings across the range of the latent variable is illustrated in the first display in Figure 29 Panel A. In contrast, the second display suggests discrepancies between model expectations and observed ratings for students at the low and high ends of the logit scale. This finding suggests that the low and high rating scale categories may not be functioning as intended by the PC model.

B. Outfit *MSE* statistics for categories are near their expected value (Outfit *MSE* ≅ 1.00). Next, quantitative indices of model-data fit to the PC model can be examined for rating scale categories. This study examines model-data fit for rating scale categories using a fit statistic that is calculated in the Facets program (Linacre, 2010): Outfit Mean Square Error (*MSE*). Outfit *MSE* is calculated by summing standardized residual variance across the observations within a rating scale category. Stated mathematically, Outfit *MSE* for rating scale categories is:

$$U_i = \frac{\sum_{n}^{N} Z_{ni}^2}{N_{ik}} \quad , \qquad (60)$$

where

$Z^2{}_{ni}$ = the standardized residual between the observed ratings and expected rating for Person *n* who receives a rating in category *k* on item *i,* based on the PC model,

and

N_{ik} = the number of persons who receive a rating in category *k*,

Because it is unweighted, the Outfit *MSE* statistic is useful because it is particularly sensitive to “outliers,” or extreme unexpected observations. As discussed in Chapter Three, limitations of Rasch fit statistics have been noted in previous research (e.g., Karabatsos, 2000; Smith, Schumacker, and Bush, 2000); however, useful applications of

Outfit *MSE* statistics have been demonstrated in the context of rater-mediated assessments (Engelhard, 2013; Linacre, 1994). Because the exact sampling distribution for *MSE* statistics is not known (Wright and Masters, 1982; Engelhard, 2013), various rules of thumb have been proposed for interpreting their values. Values that are lower than about 0.80 suggest possible dependencies among ratings. Values that are higher than about 1.20 suggest haphazard, or "noisy" ratings; extreme values in both directions warrant further investigation. Figure 29, Panel B illustrates Outfit *MSE* statistics for a four-category rating scale item. In the first set of fit statistics, all four rating scale categories meet the expectations of the PC model (Outfit *MSE* = 1.00); in the second set, only the second rating scale category displays good fit to the model.

Mokken-based Evidence of Model-Data Fit

Next, Mokken scale analysis can be used to examine model-data fit as an indicator of rating scale category effectiveness. The first indicator of model-data fit from the perspective of nonparametric IRT is the scalability coefficient based on the MH-R model. Second, indices of invariant item ordering principle from the DM-R model provide evidence of rating scale category effectiveness.

C. Item scalability coefficients (H_i) suggest scalable items (> ~ 0.3). The first indicator of model-data fit based on Mokken scale analysis is item scalability. When the MH-R model is applied to polytomous ratings, scalability coefficients describe the degree to which a set of rating scale categories form a meaningful scale (i.e., do not exhibit Guttman errors) that can be used to describe differences among students in terms of a latent variable. Mokken (1971) proposed a minimum value of H_i = 0.30 to identify items that contribute to a meaningful ordering of persons in terms of a latent variable. Although scalability coefficients do not provide diagnostic information about rating scale items in terms of individual rating scale categories, low scalability coefficients can be used to "flag" items for further investigation. Figure 29, Panel C displays H_i coefficients for a rating scale item that satisfies Mokken's (1971) criteria ($H_i = 0.56$) and a rating scale item that does not satisfy the criteria ($H_i = 0.24$).

D. Category response functions do not intersect *across items.* The next indicator of rating scale category effectiveness is based on the DM-R model requirement of nonintersecting category response functions across items. When cumulative category probabilities are considered for multiple items, non-overlapping cumulative probabilities *across* items suggest that the rating scale categories for a set of polytomous items have a meaningful order in terms of the latent variable. This requirement is based on a view of polytomous rating scales as a series of dichotomous "steps." When cumulative probabilities do not intersect, these dichotomous steps have the same order across all levels of the raw score scale, such that the interpretation of rating scale category difficulty does not depend on the student's location on the latent variable. Figure 29, Panel D illustrates two pairs of items with four-category rating scales (0 = *low*; 3 = *high*). In the first display, each of the rating scale categories for solid-line item are more difficult (smaller cumulative probability) than the corresponding categories for the dashed-line item; this pair of items meets the requirement of nonintersecting category response functions. On the other hand, the category response functions for the dashed item intersect with those for the solid item for the middle two items; this pair of items does not meet the requirement for nonintersecting category response functions. The *mokken* package (van der Ark, 2013) does not provide significance tests for violations of nonintersection at the level of rating scale categories; instead, the significance of nonintersection is examined at the overall item level.

E. Manifest invariant item ordering is observed. The next indicator of rating scale category effectiveness is based on the concept of manifest invariant item ordering (Ligtvoet et al., 2010, 2011). Manifest invariant item ordering provides evidence for rating scale effectiveness because it indicates that the overall difficulty of a set of rating scale items can be interpreted in the same way across the raw score scale which is used as a proxy for the latent variable. Figure 29, Panel E illustrates the MIIO method for examining whether expected ratings for two rating scale items overlap. In the first display, the solid item is more difficult (lower expected ratings) than the dashed item for all restscore groups. On the other hand, the difficulty ordering for the pair of items in the second display cannot be interpreted consistently across the raw score scale. Using the *mokken* package (van der Ark, 2013), nonparametric *t*-tests can be used to determine the significance of violations of MIIO. For example, if the overall average ratings on item *i* and item *j* can be ordered such that $\overline{X_i} < \overline{X_j}$, a violation of this ordering is observed for a particular restscore group *r* when this ordering is reversed, such that $(\overline{X_i} \mid R = r) > (\overline{X_j} \mid R = r)$. The significance of this violation can be examined by testing the null hypothesis that the two conditional item means are equal: $(\overline{X_i} \mid R = r) = (\overline{X_j} \mid R = r)$, against the alternative hypothesis of the reversed ordering, which is a violation of MIIO.

Results

This study explored rating scale effectiveness for the ECR items on the Alaska HSGQ Exam using the parametric PC model based on Rasch measurement theory (Masters, 1982; Wright and Masters, 1982), and adaptations of Mokken's (1971) nonparametric

MH, and DM models for use with polytomous ratings. Overall, findings from the parametric analyses indicated adequate fit to the PC model for the ECR items (Outfit *MSE* = 0.96) and the students (Outfit *MSE* = 0.96) examined in this study (see Table 16). Further, nonparametric scalability coefficients based on Mokken scale analysis suggest that the set of CR items form a medium Mokken scale (*H* = 0.58, *SE* = 0.01).

Rating Scale Guidelines

In order to address the research questions for this study, parametric and nonparametric indicators of adherence to the rating scale guidelines were examined for each of the ECR items three times: First for the overall group of students, second within the female and male subgroups, and third within the Alaskan native and white subgroups (*N* = 4,502). In this section, results are described separately for the three major rating scale guidelines: 1) Directional orientation with the latent variable, 2) Category precision, and 3) Model-data fit. Then, the overall quality of the rating scales for the ECR items is considered in terms of the diagnostic information provided by the parametric and nonparametric analyses.

1. Directional Orientation with the Latent Variable

The first guideline for rating scale effectiveness adapted from Linacre (1999, 2004) is directional orientation with the latent variable. Adherence to this guideline within the context of a rater-mediated assessment suggests that increasing locations on the latent variable (θ) correspond to ratings in higher categories. For the extended CR items examined in this study, no major violations of this guideline were observed based on parametric and nonparametric evidence. Further, no differences were observed related to this guideline for the gender or race/ethnicity subgroups.

2. Category Precision

The next guideline for rating scale effectiveness based on Linacre (1999, 2004) is related to precision with which rating scale categories describe substantively meaningful differences about students in terms of a construct. In order to investigate this guideline within the context of the HSGQ Exam, four indicators based on the PC model and two indicators based on Mokken scaling were applied to the rating data for the four ECR items and within student subgroups; results from these analyses are summarized in Table 17.

The first Rasch-based indicator of category precision is a normal or uniform distribution of ratings across categories. As can be seen in Table 17, Column A, the only item that met this requirement is ECR2. Violations of this guideline did not vary across student subgroups or across the three remaining ECR items. Specifically, the ratings formed a left-skewed distribution for ECR1, ECR3, and ECR4, with most of the ratings in Category 2. Next, the ECR ratings were examined for evidence of category precision in terms of the difference between category coefficient locations on the logit scale (Indicator B). As can be seen in Table 17, Column B, the only violation of this guideline occurred for students in the female subgroup on ECR2. Specifically, ratings assigned to female students on ECR2 violated Linacre's rule of thumb of an appropriate logit-scale difference between rating scale categories of approximately 1.40 logits to 5.00 logits, with an observed difference of 7.13 logits between the first two rating scale categories. No violations were observed for this guideline in terms of the third Rasch indicator, which required multimodal category probability functions. For the Rasch-based indicator related to evenly spaced conditional probability curves along the logit scale (Indicator D), the same violation was observed across subgroups for all of the items except ECR2. Specifically, there was a substantial "gap" between the logit-scale locations of the Rasch-Andrich thresholds for the first two categories on these items. Next, statistical information was considered for the ECR items at the overall item level (Indicator E), and in terms of individual rating scale categories (Indicator F). As can be seen in Table 17, these indices suggested threats to rating scale effectiveness for all four ECR items within at least one subgroup. For the total group of students, the overall item information function amd category information functions indicated a reduction in information related to Category 2 for all of the items except for ECR2. When these indices were examined within student subgroups, similar trends were observed. However, both the overall item information function and the category information functions suggest reduced information for students in the Alaskan native subgroup.

Indicators of category precision based on Mokken scale analysis also suggest violations of this guideline for rating scale effectiveness. First, cumulative category probabilities were examined within the four ECR items on the HSGQ Exam (Indicator G) for evidence of overlapping category response functions. When category response functions are examined *within* an item, overlapping cumulative category probabilities suggest that rating scale categories may provide redundant information about students in terms of the latent variable. As shown in Table 17, Column G, no threats to rating scale effectiveness were observed using this nonparametric indicator for the overall student group. However, violations were observed when this indicator was considered separately for students in the female and male subgroups. Specifically, category response functions related to the first two rating scale categories overlapped for students in the female subgroup whose restscores were higher than *R* = 14, and for students in the male subgroup whose restscores were higher than *R* = 15. The next Mokken-based indicator of category precision is related to the DM-R model. When items are considered in pairs, non-overlapping cumulative category probabilities *across* items suggest that the two items provide distinct information about students in terms of the latent variable. On the other hand, overlapping category response functions across items suggest that the rating scale categories do not provide unique information about a group of students. As can be seen in Table 17, Column H, overlapping cumulative category probabilities were observed for all four ECR items for the overall student sample and within the gender and race/ethnicity subgroups. Interestingly, this nonparametric indicator of rating scale effectiveness did not detect differences related to rating scale quality for student subgroups.

3. Model-Data Fit

The third guideline for rating scale effectiveness adapted from Linacre's (1999, 2004) criteria for rating scales is model-data fit for rating scale categories. Evidence for adherence to this guideline in rating data suggests that raters are applying a rating scale in a manner that is consistent with the expectations of a particular model. For this study, model-data fit for the HSGQ Exam items was considered using the PC model from Rasch measurement theory, and adaptations of Mokken's (1971) MH and DM models for use with rating data. Results from analyses related to this guideline are summarized in Table 18 for the total group of students and for students in the gender and race/ethnicity subgroups.

Inspection of Table 18 reveals that there were no violations of the model-data fit guideline for the overall group of students examined in this study based on both the parametric and nonparametric approaches. However, some misfit was observed for rating scale categories when the Rasch and Mokken models were applied separately within subgroups. First, the visual display that compares observed ratings with their expected values based on the PC model (Indicator A) suggested some unexpected ratings for students in the male subgroup with estimated latent variable locations around 5.00 logits, and for students in the white subgroup with estimated locations around 6.00 logits. Despite these unexpected observations, the Outfit *MSE* statistic for rating scale categories (Indicator B) did not detect model-data misfit for rating scale categories within student subgroups.

Next, indicators of model-data fit for rating scale categories were considered using methods from Mokken scale analysis (Mokken, 1971). The nonparametric scalability coefficient (Indicator C) did not detect misfit to MH-R model for any of the ECR items or within the gender and race/ethnicity subgroups. In contrast, the indicator of model-data fit based on the DM-R model (Indicator D), suggested model-data misfit for ECR1, ECR3, and ECR4 within some student subgroups. For ECR1, intersecting category response functions across items were observed when this indicator was considered for students in the Alaskan native and White subgroups. For ECR3, violations of the DM-R model were observed for students in the male subgroup and for students in both of the race/ethnicity subgroups. For ECR4, misfit to the DM-R model was observed for students in the male subgroup. Next, the indicator of invariant rater ordering (Ligtvoet et al., 2010, 2011) suggested violations of the model-data fit guideline within student subgroups for ECR1, ECR3, and ECR4. Invariant rater ordering was not observed for students in the male and Alaskan native subgroups for ECR1, and for students in the male, Alaskan native, and white subgroups for ECR3 and ECR4.

Correspondence between Parametric and Nonparametric Indicators

The last research question for this study focuses on the correspondence between indicators of rating scale effectiveness based on parametric and nonparametric IRT models. Overall, results from this study suggest that the parametric PC model and the nonparametric MH-R and DM-R models provided related, but slightly different, information about the overall quality of the rating scales for the ECR items on the HSGQ Exam. First, both the Rasch- and Mokken-based indicators based on the first guideline suggested directional orientation with the latent variable. Next, indicators of category precision based on Rasch measurement theory and Mokken scaling suggested violations of the second guideline related to all four ECR items for the overall group of students and within the gender and race/ethnicity subgroups. However, some differences between the parametric and nonparametric approach were observed among the indicators for this guideline. Although both the Rasch- and Mokken-based indicators revealed differences in rating scale functioning for gender subgroups on ECR2 (Indicators B and G), differences in rating scale functioning for Alaskan native students were only detected by the parametric indicators related to statistical information (Indicators E and F). Finally, the parametric and nonparametric indicators of model-data fit for rating scale categories provided slightly different conclusions about the quality of the rating scales for the ECR items within the gender and race/ethnicity subgroups. Interestingly, more violations were observed based on the nonparametric indicators of rating scale effectiveness than were identified by the parametric indicators. This finding highlights the diagnostic value of nonparametric methods as a methodological tool to identify aberrant patterns in rating scale use for rater-mediated assessments.

Conclusions

The major purpose of this chapter was to develop and illustrate a set of diagnostic indicators of rating scale effectiveness based on Mokken (1971) scaling. The indices were adapted from Linacre's (1999, 2004) guidelines for rating scales, and considered alongside parametric indicators based on the PC formulation of the Rasch model (Masters, 1982; Wright & Masters, 1982). These indices of rating scale effectiveness were considered for the overall group of students and within gender and race/ethnicity subgroups. Further, the degree to which nonparametric indicators of rating scale effectiveness based on polytomous versions of Mokken's nonparametric IRT models provided similar information to the parametric indicators based on the PC model was also explored. This section provides conclusions for the three research questions that guided the analyses.

1. What does Rasch measurement theory reveal about the quality of a rating scale for a large-scale rater-mediated writing assessment within and across student subgroups?

The first research question for this chapter focused on the use of parametric IRT to diagnose rating scale effectiveness. Because they meet the requirements for invariant measurement, models based on Rasch measurement theory (Rasch, 1960/1980) have been frequently applied as methodological tools for examining rating scale functioning in social science research (Wright & Masters, 1982). In particular, the PC formulation of the

Rasch model (Masters, 1982; Wright & Masters, 1982) is useful in this context because it allows the structure of the rating scale to be investigated separately for each item. In this study, the parametric analyses revealed that the set of extended CR items on the HSGQ Exam demonstrated adequate overall fit to the PC model for students and ECR items. This overall finding suggests that the PC model calibrations can be trusted as appropriate interval-level representations of the students, ECR items, and rating scale categories for the writing section of the HSGQ Exam in terms of the latent variable. In other words, estimates of student achievement can be described independently from the particular items that they happened to take on the Spring 2013 administration of the Alaska HSGQ Exam, and estimates of ECR item difficulty can be described independently from the particular sample of students who participated in their administration.

When the separate indicators for the three major guidelines were considered, the statistics and displays provided by the PC model revealed that each indicator provided slightly different diagnostic information about the rating scales for the ECR items. Although observed violations in this study were minor, the illustrations in this chapter emphasized the utility of the PC model for identifying violations of the guidelines which suggest that the relationship between students, items, and ratings may not be appropriately represented by the restrictive form of the PC model. As shown in this chapter, this diagnostic information highlights specific rating scale categories or ranges of student achievement where the application of the rating scale was inconsistent across ECR items or student subgroups—information that can be applied in practice to improve the development of items and scoring rubrics, or to improve rater training related to these inconsistencies.

2. What does Mokken scale analysis reveal about the quality of a rating scale for a large-scale rater-mediated writing assessment within and across student subgroups?

The second research question for this chapter focuses on the use of Mokken scale analysis (Mokken, 1971) as a method for evaluating the quality of rating scales in large-scale rater-mediated assessments. Overall findings from the application of the MH-R and DM-R models to the Alaska HSGQ Exam data suggested that the rating data formed a medium Mokken scale (H = 0.58, SE = 0.01), and that some significant violations of double monotonicity were apparent. Thus, from the perspective of Mokken scaling, the HSGQ Exam can be said to provide an item-independent ordering of students in terms of the latent variable, and that some revisions of items, rating scales, or rating scale use may be necessary before these items can provide a person-independent ordering of items in terms of the latent variable. Beyond these overall indices of model-data fit, this study proposed a new set of nonparametric criteria for evaluating the quality of rating scales that were adapted from the set of rating scale guidelines proposed by Linacre (1999, 2004). Because these indices were applied as methods for examining rating scales, the interpretation of the statistics and displays from the MH-R and DM-R models were slightly different than the rating quality displays discussed in Chapter Four. Inspection of findings from the set of nonparametric indicators suggested that each indicator provides slightly different information about the functioning of rating scale items in terms of the assumptions of the MH-R and DM-R model. Similar to the parametric approach, information about violations of the guidelines for the overall student subgroup, and within gender and race/ethnicity subgroups that is provided by these nonparametric indicators can be used to identify areas for improvement in assessment development and rater training. Because the assumptions of the nonparametric approach are less restrictive than the parametric requirements of the Rasch model, these indicators are particularly useful during the early stages of assessment development, or during rater training.

3. Do indices of rating scale effectiveness based on Rasch measurement theory and Mokken scale analysis lead to comparable conclusions about the overall quality of a rating scale?

The third research question for this chapter focuses on the degree to which parametric and nonparametric indicators of rating scale effectiveness lead to comparable conclusions about the quality of rating scales in large-scale rater-mediated writing assessment. When the set of indicators based on the PC, MH-R, and DM-R models were inspected using the empirical data, some differences were observed in the ability to diagnose violations of the three guidelines. Specifically, the nonparametric analyses revealed violations of the three guidelines that were not diagnosed by the parametric indices based on the PC model (Masters, 1982; Wright & Masters, 1982), particularly when analyses were conducted separately within student subgroups. These findings suggest that Mokken scale analysis provides useful information that can augment parametric methods for examining the requirements of invariant measurement.

In summary, the analyses by the three models offer mixed evidence regarding the functioning of the rating scales. Nonparametric methods offer promising new guidelines for examining facets of invariance that are not examined with current parametric approaches to IRT.

Chapter Six: Discussion and Conclusions

This dissertation has explored Mokken's (1971) nonparametric theory and procedure for scale analysis as a method for evaluating the quality of rater-mediated educational assessments. Rating quality indices based on Rasch measurement theory (Rasch, 1960/1980) were used as a frame of reference for considering nonparametric indicators of rating quality based on Mokken scaling. Following the theoretical framework illustrated in Figure 1, this study employed the combination of a lens model for rater-mediated assessments, a theory of invariant measurement, and selected indicators of rating quality based on measurement models with useful properties in order to evaluate the quality of rater-mediated assessments.

As a method for informing the interpretation and use of rater-assigned scores, this study proposed and illustrated diagnostic indicators that can be used to evaluate rater-mediated assessments in terms of individual raters and rating scales. These ideas were explored in depth in the dissertation using the following research questions:

1. What are the major underlying measurement issues related to rating quality?
2. How have these measurement issues been traditionally addressed in previous research?
3. How has Rasch Measurement Theory been used to examine the quality of ratings?
4. How can Mokken scaling be used to examine the quality of ratings?
5. What is the relationship between Mokken- and Rasch-based indices of rating quality?

First, previous research related to rater-mediated assessments was examined in order to identify persistent concerns and traditional quality-control indices in rater-mediated assessments (see Chapter One and Chapter Two). Then, modern methods for monitoring the quality of ratings based on Rasch measurement theory (Rasch, 1960/1980; see Chapter Three) and Mokken scaling (Mokken, 1971; see Chapter Four) were explored using illustrative data analyses. Finally, the correspondence between these two modern approaches to monitoring rating quality was considered using data from a recent administration of a large-scale statewide writing assessment (see Chapter Five).

The final chapter of this dissertation is organized in two major sections. First, each research question is considered in terms of the findings from the dissertation study. The second section discusses limitations of the study, implications for research, policy, and practice, and directions for future research.

Research Question 1: What are the major underlying measurement issues related to rating quality?

In this study, rater-mediated assessments were defined as procedures in which a rater judges the quality of a student's response to a task using one or more domains defined by a rubric that is designed to represent a construct using a rating scale. Motivated by the concept of "washback," or the pedagogical implications for different types of high-stakes assessments, the development and use of rater-mediated assessments in large-scale settings has been accompanied by concerns related to the reliability, validity, and fairness of rater-assigned scores (Lane & Stone, 2006). An aspect of validity, indicators of rating quality provide information about the meaning of rater-assigned scores as reflections of a construct (Messick, 1995).

In this study, rating quality was considered using a theory of human judgment based on a lens model that illustrated the complex nature of rater-mediated assessments. Specifically, a lens model for rater-mediated assessment (Figure 3) was presented in

Chapter One based on Brunswick's (1952) lens model (Figure 2) and Social Judgment Theory (Cooksey, 1996a; Hammond, Stewart, Brehmer, & Steinmann, 1975) that highlighted the influence of various cues, or mediating variables, on rater judgment. The lens model highlights several important aspects of rater-mediated assessments. Beginning with the focal variable, θ_P, the model emphasizes the fact that a student's "true" score in terms of the latent variable is an unobserved and unobservable variable in the context of educational achievement tests. In terms of the distal variable, θ_R, the model highlights observed ratings as *judged* locations of a student in terms of a construct that are influenced by various cues within complex ecological contests. The lens model for rater-mediated assessment emphasizes the fact that a rater's judgment (θ_R) about a student's location on the construct (θ_P) is informed by cues such as domains on an analytic rubric, student benchmark performances that represent levels of achievement specific to a particular assessment, and the rating scale that corresponds to the rubric for a particular assessment. The types and role of these cues are context-specific, and each rater-mediated assessment system must be considered in terms of its unique ecological context. Because human raters exist within ecological contexts that mediate the relationship between θ_P and θ_R, the major underlying measurement issues related to rating quality include concerns about a rater's ability to provide a "clear reflection" of a student's performance within a particular assessment context.

In Chapter Two, previous research on rater-mediated assessments was examined in order to identify persistent measurement concerns related to rating quality. Beginning with Edgeworth's (1890) experiments on the consistency of ratings in psychophysical experiments and in rating the quality of compositions, the review of research suggested a common theme in research over the last century of concerns related to the seemingly subjective nature of human judgment. In terms of the lens model, research on the quality of rater-assigned scores reflects a view of consistency as evidence that mediating variables do not unduly influence rater interpretation of a performance. Recent discussions of concerns with the quality of ratings suggested that these concerns remain prevalent in research on rater-mediated assessments (Hamp-Lyons, 2011; Wolfe & McVay, 2012).

In summary, the main findings related to the first research question for this study are that the major underlying measurement issues related to rating quality stem from concerns with the influence of mediating variables on the interpretation and use of rater-assigned scores. The continued focus on evaluating the quality of ratings in current research suggests that these concerns are persistent in educational assessment. Further, these persistent concerns with the influence of human judgment on descriptions of student achievement emphasize the need for continued research on the use of quality control indices at various stages of the development and implementation of rater-mediated assessment systems, along with continued exploration of complex ecological contexts that surround and define rater-mediated assessments.

Research Question 2: How have these measurement issues been traditionally addressed in previous research?

The second research question for this study focused on methods with which the persistent concerns related to rating quality have been addressed in previous research. This research question was addressed in Chapter Two using a review of literature on traditional methods for detecting and describing rating quality. In terms of the lens model, the research that was reviewed in Chapter Two represents traditional attempts at evaluating the match between θ_P and θ_R using evidence of rating quality based on statistical indicators.

Since Thorndike's (1920) description of halo errors, and Guilford's (1936) system for identifying rater leniency and severity errors, halo errors, or indices of interrater reliability and agreement, a variety of methods have been developed to monitor rating quality during rater training and operational scoring. The three major categories of traditional rating quality indices are 1) rater agreement, 2) rater errors and systematic biases, and 3) rater accuracy (Murphy & Cleveland, 1991; Johnson, Penny, & Gordon, 2009). Rater agreement indicators describe the degree to which raters assign matching scores to the same performance. Indices of rater errors and systematic biases describe specific patterns or trends in rating behavior that are believed to contribute to the assignment of scores different from those warranted by a student's performance. Finally, rater accuracy indicators describe the match between operational ratings and those established as "true" or "known" ratings by individuals or committees of expert raters. When indices of rating quality based on these three categories are applied, high levels of agreement, low levels of error and systematic bias, and high levels of accuracy are assumed to reflect high-quality ratings. A variety of indices of rating quality have been developed within these three categories; an overview of the most-commonly applied indices within each category was provided in Chapter Two. Despite the development of rating quality indices based on modern measurement models, such as those described in this dissertation, most operational methods for monitoring rating quality continue to use indices based on these three categories (Johnson, Penny, & Gordon, 2009; Murphy & Cleveland, 1991).

In summary, the main findings for this research question suggest that traditional methods for evaluating rating quality focus on a disparate set of statistical summaries of rating patterns that are assumed to reflect evidence of rater agreement, errors and systematic biases, and accuracy. This traditional approach to monitoring rating quality remains prevalent in most large-scale rater-mediated assessment systems.

Research Question 3: How has Rasch measurement theory been used to examine the quality of ratings?

The third research question for this study focused on a modern approach to monitoring rating quality based on Rasch measurement theory (Rasch, 1960/1980). The use of statistics and displays based on Rasch models as indicators of rating quality is not a new idea. Several scholars have proposed the use of Rasch-based rating quality indices a method for examining rating quality in terms of rater-invariant measurement (Eckes, 2011; Engelhard, 2013; Myford & Wolfe, 2003, 2004), and in terms of rater accuracy (Engelhard, 1996, 2013; Wind & Engelhard, 2012, 2013). Further, the relationship between Rasch-based indices of rater invariance and rater accuracy suggests that Rasch-based fit statistics for observed ratings may provide information related to direct indicators of rater accuracy for overall and domain-level ratings (Wind & Engelhard, 2012, 2013). In Chapter Three, previous applications of Rasch measurement theory as a methodological toolkit for exploring rating quality were summarized, and rating quality indices based on this approach were illustrated using an example dataset. The statistics and displays for monitoring ratings presented in Chapter Three, along with the statistics and displays for examining rating scale effectiveness in Chapter Five, highlight the versatility of Rasch-based rating quality indices for different rating designs and assessment systems. For example, the Rasch-based rating quality indices that were presented in this dissertation can be used to examine the requirements of rater-invariant measurement in the case of holistic or analytic ratings, complete or incomplete rating designs, and in the case of the combination of rating scales with different numbers of categories. These rating quality indices are promising as a method for empirically investigating rater invariance and rater accuracy during rater training and operational scoring. However, confidence in the inferences that are drawn from Rasch-based calibrations of raters, students, items, domains, and other aspects of an assessment system depends on confidence in the tenability of the parametric requirements that underlie Rasch models.

In summary, the main findings related to Research Question 3 are that the parametric IRT models within the framework of Rasch measurement theory provide a useful set of statistical and graphical summaries that can be used to evaluate a set of ratings for evidence of rater-invariant measurement. These rating quality indices differ from the indicators based on rater agreement, error and systematic biases, and accuracy described in Chapter Two. Whereas the traditional approach focuses on replications of "ideal" raters who agree with one another, demonstrate a lack of errors and systematic biases, and match the ratings assigned by experts, Rasch models for ratings are probabilistic, such that variance in ratings is necessary in order to construct measures from ratings that describe students, raters, and other tasks in terms of a latent variable.

Research Question 4: How can Mokken scaling be used to examine the quality of ratings?

Based on the idea that the functional form requirements of parametric Item Response Theory (IRT) models, such as the Rasch model, are not warranted in social science and behavioral applications, nonparametric models have been proposed as an alternative method to examine item response data. In this study, Mokken's (1971) nonparametric IRT models were adapted for use with rater-assigned polytomous scores as a method for evaluating rater-mediated assessments. The nonparametric, or ordinal, approach to measurement that characterizes Mokken's procedure for scale analysis is desirable in settings where relations among variables are difficult to define, such as rater perceptions of student achievement. Among the major motivations for the application of nonparametric IRT models in the social and behavioral sciences is the lack of confidence in the assumption that transforming ordinal observations (such as ratings) to an interval scale is an appropriate way to describe a latent construct, which may or may not possess these interval-level properties. In other words, there is a distinction between data analyses with transformed observations assumed to reflect a construct and the actual properties of the construct. Further, as pointed out by Cliff and Keats (2003), the desired conclusions to be drawn from these investigations are usually ordinal in nature, and often do not require the interval-level metric that is achieved through the application of parametric models. Mokken's (1971) approach to scaling provides a method for examining the degree to which a set of observations adhere to important aspects of measurement, such as monotonicity and non-intersection, without imposing potentially inappropriate assumptions on the level of measurement.

In this study, Molenaar's (1982, 1997) and Sijtsma and Hemker's (1992) polytomous adaptations of Mokken's original MH and DM models were explored as methods for monitoring the quality of ratings in large-scale rater-mediated assessments. Parallel to the presentation of parametric indicators of rating quality based on Rasch measurement theory, a set of nonparametric indices of rater effects were illustrated in

Chapter Four, and nonparametric indicators of rating scale effectiveness were explored in Chapter Five. This study found that statistics and displays based on Mokken scaling provide information that can be used to evaluate rater-mediated assessments in terms of a variety of desirable properties, including monotonicity, scalability, and nonintersecting response functions. Because the DM model is based on invariant ordering properties, these indicators of rating quality can be used to investigate invariance in rater-mediated assessments related to student and rater ordering in terms of a construct.

In summary, the major findings related to the fourth research question suggest that Mokken scale analysis provides a method for examining the degree to which a set of ratings adheres to a set of underlying assumptions based on the principles of invariance that are not as strict as the requirements of parametric IRT models. Overall, the illustrative data analyses suggested that Mokken scale analysis provides useful information that can augment parametric methods for examining the requirements of invariant measurement, such as those provided by the Rasch model (Rasch, 1960/1980).

Research Question 5: What is the relationship between Rasch- and Mokken-based indices of rating quality?

The final research question for this study explores the relationship between rating quality indices based on a parametric (Rasch) and nonparametric (Mokken) approach to IRT. Chapter Three and Chapter Four explored rating quality indicators based on these two approaches separately; these chapters revealed the implications of the unique properties of measurement models based on each approach for evaluating rater-mediated assessments. In Chapter Five, a slightly different view of measurement quality in rater-mediated assessments was considered, and Mokken-based quality control indices were explored using Rasch-based indices as a frame of reference. In order to supplement the indicators of rating quality that focus on individual rater effects, Chapter Five presented a set of techniques that can be used to empirically investigate the structure of rating scales in terms of how individual rating categories are applied. These indices were classified under three major guidelines for rating scale effectiveness: 1) directional orientation with the latent variable, 2) category precision, and 3) model-data fit. Using data from the Alaska High School Graduation Qualifying exam, the three guidelines were examined for the overall group and within gender and race/ethnicity subgroups.

Results from the empirical analyses indicated that the Rasch-based PC model and (Masters, 1982; Wright & Masters, 1982) and polytomous Mokken scaling (Mokken, 1971) provide related, but slightly different, information about the structure of a rating scale. Interestingly, more violations of the guidelines for rating scale effectiveness were observed when nonparametric indicators were considered than were identified by the parametric indicators. This finding highlighted the diagnostic value of nonparametric methods as a methodological tool to identify aberrant patterns in rating scale use for rater-mediated assessments.

In summary, the major findings from this study for the final research question suggest that the PC model, the MH-R model, and the DM-R model for polytomous ratings offer mixed evidence regarding measurement quality when they are used to evaluate the functioning of rating scales. Nonparametric methods offer promising new guidelines for examining facets of invariance that are not examined with current parametric approaches to IRT.

Limitations

There are several limitations that should be considered when drawing inferences for research, theory and practice based on this dissertation. The first limitation of this study is related to the generalizability of results. This study used secondary data from two large-scale, K-12, writing assessments that included essay or constructed-response components. As a result, statistical generalizations to populations and rater-mediated assessments in other subject areas and at other grade levels should be made with caution based on this research. Further, the use of secondary data prevented researcher control of the rating designs and the variables about which data were collected. Recognizing these limitations, this study emphasized the illustration of methodological techniques for evaluating rater-mediated assessments rather than emphasizing conclusions about particular assessments, subject areas, grade levels, or rating designs. Furthermore, the relationship between parametric and nonparametric models and their associated rating quality indices is likely to be generalizable beyond the particular contexts examined in this study.

Another limitation is that this study did not attempt to investigate all possible parametric and nonparametric methods for monitoring rating quality. Instead, the intent of the dissertation was to highlight two key approaches to evaluating rater-mediated assessments: parametric models based on Rasch measurement theory (Rasch, 1960/1980) and nonparametric models based on Mokken scale analysis (Mokken, 1971). The use of different models within each approach, such as the parametric graded response model (Samejima, 1997) and methods based on Cliff and Keats' (2003) ordinal test theory, may lead to different results and conclusions.

Implications for Research, Theory, Policy, and Practice

In this section, the importance of this dissertation for research and theory in the area of rater-mediated assessments is examined. Then, implications for policy and practice are considered. Finally, directions for future research using nonparametric Item Response Theory as a method for evaluating rater-mediated assessments are discussed.

Research and Theory

The most significant implications of this study are in the areas of research and theory related to rater-mediated assessments. This study highlighted the complex nature of rater-mediated assessment systems as ecological contexts in which rater-assigned scores may be influenced by a variety of mediating variables. These mediating variables may cloud the interpretability of ratings as accurate reflections of a construct. This view of rater-mediated assessments brings to mind Mokken's (1971) reservations related to the functional form requirements that underlie parametric IRT models in social and behavioral research. Although Rasch recognized the invariant ordering properties achievable through the use of raw scores, he argued that this method for describing persons, items, and constructs does not result in measurement:

> It seems reasonable to state *that an ordering of persons by [raw score] is an ordering by their ability to solve this type of problem….* This *ordering* of items and persons, however, does not imply a *measurement* of degrees of difficulty and ability on a ratio scale. (Rasch, 1960/1980, p. 69, italics in original)

While Mokken (1971) did not reject this perspective, he argued that the application of parametric models might lead to inappropriate conclusions about variables that are not clearly understood:

> In vast areas of social research the application of parametric models may often be too far fetched. Their application presupposes a relatively deep insight into the structure of the variable to be measured and the properties of the items by which it can be measured (Mokken, 1971, p. 173)

By examining the relationship between a parametric and nonparametric approach to IRT within the context of rater-mediated assessments, this study extends previous research on nonparametric IRT in general by applying nonparametric models to a new context. This comparison sheds light on the relationship between Mokken's (1971) nonparametric IRT models and parametric IRT models based on Rasch measurement theory that can be used to inform the interpretation of both approaches separately, and to inform future comparisons using these and other parametric and nonparametric models.

Second, this research extends previous methods for monitoring rating quality by proposing and illustrating a set of diagnostic indicators of rating quality based on Mokken scale analysis. Because Mokken's (1971) nonparametric IRT models are based on invariant ordering principles, the rating quality indices presented in this study provide researchers with additional tools for understanding and evaluating rater-invariant measurement. The requirements for rater-invariant measurement were presented in Chapter One:

1. The measurement of persons must be independent of the particular raters that happen to be used for the measuring: *Rater-invariant measurement of persons*.
2. The calibration of the domains must be independent of the particular raters used for calibration: *Rater-invariant calibration of domains.*

3. The structure of the rating categories must be independent of the particular raters used for calibration: *Rater-invariant calibration of rating scales*.
4. Persons, raters, domains, and rating categories must be simultaneously located on a single underlying latent variable: *Variable map.*

This study used illustrative and exploratory data analyses to investigate the use of Mokken's (1971) nonparametric models as a methodological approach to evaluating these requirements in empirical rating data. Additional research is needed in order to more fully understand the relationship between Mokken-based rating quality indicators and conclusions about rater-invariant measurement (discussed below). However, several important observations were made through this initial application:

- This is the first application of Mokken (1971) scaling to rater-mediated educational assessments.
- This research provided an empirical example of the utility of Mokken scaling for identifying departures from rater-invariant measurement.
- It is important to consider the tenability of the particular requirements or assumptions that underlie a measurement model before interpreting rating quality indices based on the model.

Policy and Practice

This study also has implications for policy and practice. The overall goal in developing indicators of rating quality is to provide a method for evaluating the quality of rater-mediated assessments that can inform score interpretation and use for large-scale rater-mediated assessments. The indicators of rating quality proposed and illustrated in this dissertation are most applicable during the assessment development and rater training stages of rater-mediated assessments. In particular, this study highlighted the need to incorporate quality control procedures that go beyond the traditional indices of rater agreement, error, and accuracy, and incorporate the requirements of rater-invariant measurement, especially in high-stakes contexts such as the next-generation assessments that are currently being implemented as part of the Common Core State Standards initiative. Although additional research is needed in order to more fully understand the utility of nonparametric indicators of rating quality, results from this study suggest that these statistics and displays can be used as supplementary information for examining invariance in rater-mediated assessment systems.

Future Research

Future research is needed in order to develop a more complete understanding of the application of Mokken scaling to rater-mediated educational assessments. Specifically, three major areas for future research are of note: 1) theoretical and empirical comparisons of parametric and nonparametric models for rater-mediated assessments, 2) challenges associated with the application of nonparametric models to different designs of rater-mediated assessments, and 3) the utility of nonparametric rating quality indices in practice for improving the quality of ratings.

First, additional research is needed that includes conceptual and theoretical comparisons of parametric and nonparametric models within the context of rater-mediated assessments. In addition to the Rasch- and Mokken-based models presented in this study, comparisons with other parametric and nonparametric models will inform the choice of an appropriate method for describing and monitoring measurement quality in these contexts. Second, challenges in applying Mokken scaling techniques to rater-mediated assessments need to be addressed. Specifically, the application of the Monotone Homogeneity for Ratings model and Double Monotonicity for Ratings model in cases of incomplete rating designs, analytic scoring, and the use of items with different numbers of rating scale categories is not straightforward. Resolution of these issues is essential to the widespread application of nonparametric IRT-based methods to monitor rating quality. Finally, research is needed that considers the practical utility of the Mokken-based rating quality indices presented in this dissertation. Specifically, a complete understanding of the implications for this study requires investigation of the degree to which nonparametric IRT indices can appropriately identify raters in need of remediation during rater training and operational scoring in a large-scale assessment setting.

Overall, the nonparametric rating quality indices developed in this study provide an exploratory approach to examining the psychometric quality of rater-assigned scores that can inform research, theory, and practice related to rater-mediated assessments.

References

Abedi, J. (1996). Interrater/test reliability system (ITRS). *Multivariate Behavioral Research*, *31*(4), 409-417.

Agresti, A. (1992). Modeling patterns of agreement and disagreement. *Statistical Methods in Medical Research*, *1*(2), 201-218.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (in preparation). *Standards for educational and psychological testing*. Washington, DC: AERA.

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

Andrich, D. A. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), 7-16.

Andrich, D. A. (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.

Andrich, D. A., de Jong, J. H. A. L., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. R. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 58-68). Munster, Germany: Waxmann-Verlag.

Banerjee, M. (2006). Interrater agreement. In S. Kotz, C. B. Read, N. Balakrishnan, & B. Vidakovic (Eds.), *Encyclopedia of statistical science* (Volume 6, Second Edition) (pp. 3,619–3,626). New Jersey: John Wiley & Sons.

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3-23.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, *28*(1), 51-75.

Bech, P., Hansen, H. V., & Kessing, L. V. (2006). The internalising and externalising dimensions of affective symptoms in depressed (unipolar) and bipolar patients. *Psychotherapy and Psychosomatics*, *75*(6), 362-369.

Bennett, R. E. (1993). On the meaning of constructed response. In R. E. Bennett and W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 1-27). Hillsdale, NJ: Erlbaum.

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

Bingham, W. V. (1939). Halo, valid and invalid. *Journal of Applied Psychology, 23,* 221-228.

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.

Brennan, R. L. (1995). Standard setting from the perspective of Generalizability theory. In M.L. Bourque (Ed.), *Joint Conference on Standard Setting for Large-Scale Assessments* (pp. 269-287). Washington, DC: NCSE-NAGB.

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (NCES 96-802) (pp. 198-258). Washington, DC: National Center for Education Statistics.

Brennan, R. L. (2000). Performance assessments from the perspective of Generalizability theory. *Applied Psychological Measurement*, *24*(4), 339-353.

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

Brehmer, B. & Joyce, C. R. B. (Eds.) (1988). *Human judgment: The SJT view.* Amsterdam: North-Holland Elsevier.

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

Brunswik, E. (1957). Scope and aspects of the cognitive problem. In H. E. Gruber, K. R. Hammond, & R. Jessor (Eds.), *Contemporary approaches to cognition: A symposium held at the university of Colorado* (pp. 5-31). Cambridge: Harvard University Press.

Burry-Stock, J. A., Shaw, D. G., Laurie, C., & Chissom, B. S. (1996). Rater agreement indexes for performance assessment. *Educational and Psychological Measurement*, *56*(2), 251-262.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, *24*(4), 310-324.

Clauser, B. E., Clyman, S. G., & Swanson, D. B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement*, *36*(1), 29-45.

Cliff, N. & Keats, J. A. (2003). *Ordinal measurement in the behavioral sciences.* Mahwah: Lawrence Erlbaum Associates.

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213.

Cooksey, R. W. (1996a). The methodology of social judgment theory. *Thinking & Reasoning*, *2*(2-3), 141-174.

Cooksey, R. W. (1996b). *Judgment analysis: Theory, methods, and applications*. San Diego: Academic Press.

Cooksey, R. W., Freebody, P., & Bennett, A. J. (1990). The ecology of spelling: A lens model analysis of spelling errors and student judgments of spelling difficulty. *Reading Psychology: An International Quarterly*, *11*(4), 293-322.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90,21-* 244.

Cooper, A., Levin, B., & Campbell, C. (2009). The growing (but still limited) importance of evidence in education policy and practice. *Journal of Educational Change, 10*, 159-171.

Cole, N. S. & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-220). New York: American Council on Education and Macmillan.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37(2), 163-178.*

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

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

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

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671-684.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist*, *30*(1), 1-14.

Cronbach, L. J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373-399.

de Ayala, R. J. (2009) *The theory and practice of item response theory.* New York: Guilford Press.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, *42*(1), 53-76.

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

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

Doherty, M. E. (Ed.) (1996). *Social Judgment Theory* (special issue of *Thinking and Reasoning, 2*(2/3), 105-248). East Sussex, UK: Psychology Press.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, *14*(2), 88-115.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*(3), 197-221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H). Strasbourg, France: Council of Europe/Language Policy Division.

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

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society*, *53*(3), 460-75.

Educational Testing Service (2010). *TOEFL iBT test scores.* Retrieved from: http://ets.org/toefl/ibt/scores/

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*,*2*(3), 175-196.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. New York: Peter Lang.

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

Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement, 1*(1). 19-33.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. Haladyna (Eds.), *Large-scale assessment programs for ALL students: Development, implementation, and analysis,* (pp. 261-287). Mahwah, NJ: Erlbaum.

Engelhard, G., Jr. (2005). Item response theory (IRT) models for rating scale data. In B. S. Everitt and D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral sciences, Vol. 2* (pp. 995-1003). Chichester: John Wiley & Sons, Ltd.

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

Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research & Perspective*, *6*(3), 155-189.

Engelhard, G., Jr. (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., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences.* New York: Routledge.

Engelhard, G. Jr., & Myford, C. M. (2003). Monitoring faculty consultant performance in the Advanced Placement English Literature and Composition Program with a many-faceted Rasch model. (College Board Research Report No. 2003-1). New York: College Entrance Examination Board.

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

Engelhard, G., Jr., Wind, S. A., Kobrin, J., & Chajewski, M. (in press). *Differential Item and Person Functioning in Large-scale Writing Assessments within the Context of the SAT Reasoning Test.* College Board Research and Development Report.

Fechner, G. T. (1860). Elements of psychophysics. In W. Dennis (Ed.). *Readings in the History of psychology*. New York: Applenton-Century-Crofts, 1948.

Fisher, R. A. (1958). *Statistical methods for research workers*. New York: Hafner Publishing Co.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382.

Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613-619.

Goldstein, W. M. (2004). Social judgment theory: Applying and extending Brunswik's probabilistic functionalism. In D. J. Koehler & N. Harvey (Eds.) *Blackwell handbook of judgment and decision-making* (pp. 37-61). Malden, MA: Blackwell.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross-classifications. *Journal of the American Statistical Association*, *49*(268), 732-764.

Guilford, J. P. (1936). *Psychometric methods.* New York: McGraw-Hill.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10* (4), 255-282.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, and S. A. Clausen (Eds.), *Measurement and Prediction* (Volume IV, pp. 60-90). Princeton: Princeton University Press.

Gyagenda, I. S., & 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.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Applied Measurement in Education, 12*(3), 38-47.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, *62*(4), 255.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* New York: Oxford University Press.

Hammond, K. R. & Joyce, C. R. B. (eds.) (1975) *Psychoactive drugs and social judgment: Theory and research*. New York: Wiley.

Hammond, K. R., & Stewart, T. R., (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications.* Oxford: Oxford University Press.

Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. (1975). Social judgment theory. In M. F. Kaplan and S. Schwartz (Eds.), *Human judgment and decision processes* (pp. 271-312). New York: Academic Press.

Hammond, K. R., & Wascoe, N. E. (Eds.). (1980). *Realizations of Brunswik's representative design.* San Francisco: Jossey-Baas.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing, 8*(1), 5-6.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, *12*(1), 1-9.

Hamp-Lyons, L. (2011). Writing assessment: Shifting issues, new tools, enduring questions. *Assessing Writing*, *16*(1), 3-5.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(1), 43-58.

Harman, H. H. (1976). *Modern factor analysis*. Chicago: University of Chicago Press.

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability correlation versus percentage of agreement. *Written Communication*, *16*(3), 354-367.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1) 77-89.

Hemker, B. T., Sijtsma, K., Molenaar, I. & Junker, B. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61*, 679-693.

Hieronymous, A., Hoover, H., Cantor, N., & Oberley, K. (1987). *Handbook for focused holistic scoring.* Chicago: Riverside Publishing.

Hogan, T. P., & Mishler, C. (1980). Relationships between essay tests and objective tests of language skills for elementary school students. *Journal of Educational Measurement*, *17*(3), 219-227.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? A Generalizability theory approach. *Assessing Writing*, *13*(3), 201-218.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, *23*(4), 310-324.

Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika, 59,* 77-79.

Johnson, R. L., Penny, J. & Gordon, B. (2000). The relationship between score resolution and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13*(2), 121-138.

Johnson, R. L., Penny, J., & Gordon, B. (2001). Score resolution and the interrater reliability of holistic scores in rating essays. *Written Communication*, *18*(2), 229-249.

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

Johnson, R. L., Penny, J., Fisher, S., & Kuhs, T. (2003). Score resolution: An investigation of the reliability and validity of resolved scores. *Applied Measurement in Education*, *16*(4), 299-322.

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

Jorsekog, K. G. (2007). Factor analysis and its extensions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 47-77). Mahwah, NJ: Erlbaum.

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359-1378.

Juslin, P., & Montgomery, H. (Eds.). (1999). *Judgment and decision making: Neo-Brunswikian and process-tracing approaches*. Lawrence Erlbaum.

Kaliski, P., Wind, S. A., Engelhard, G., Morgan, D., Reshetar, R., & Plake, B. (2013). Using the Many-Facet Rasch model to evaluate standard-setting judgments: Setting performance standards for Advanced Placement examinations. *Educational and Psychological Measurement, 73*(2), 1-26.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527.

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

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*(2), 152-176.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404-426.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1), 81-93.

Kingsbury, F. A. (1922). Analyzing ratings and training raters. *Journal of Personnel Research, 1,* 377-382.

Kingsbury, F. A. (1933). Psychological tests for executives. *Personnel, 9,* 121-133.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing*, *28*(2), 179-200.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(1), 26-43.

Kobrin, J. L., & Kimmel, E. W. (2006). *Test development and technical information on the Writing Section of the SAT Reasoning Test.* New York: College Board (Research Notes, RN-25)

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30,* 61—70.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (Second Edition). Princeton University Press, Princeton.

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Lane, S., & Stone, C. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational Measurement, Fourth Edition* (pp. 387-431). Westport, CT: American Council on Education and Praeger.

Lakatos, I. (1978). *The methodology of scientific research programs.* Cambridge, UK: Cambridge University Press.

Laudan, L. (1977). *Progress and its problems: Toward a theory of scientific change*. Berkeley, CA: University of California Press.

Lazarsfeld, P. (1966). Concept formation and measurement in the behavioral sciences: Some historical observations. In G. J. Direnzo (Ed.), *Concepts, theory, and explanation in the behavioral sciences* (pp. 144-202). New York: Random House.

LeBrenton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815-852.

Lehmann, E. C. (1986). *Testing statistical hypotheses (2nd Ed)*. John Wiley & Sons, New York.

Licht, R. W., Qvitzau, S., Allerup, P., & Bech, P. (2005). Validation of the Bech–Rafaelsen Melancholia Scale and the Hamilton Depression Scale in patients with major depression: Is the total score a valid measure of illness severity? *Acta Psychiatrica Scandinavica*, *111*(2), 144-149.

Ligtvoet, R., van der Ark, L. A., Bergsma, W. P. & Sijtsma, K. (2011). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika, 76,* 200-216.

Ligtvoet, R., van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*(4), 578-595.

Linacre, J. M. (1989/1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103–122.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. Smith and R. Smith (Eds.) *Introduction to Rasch measurement: Theory, models, and applications* (pp. 258-278). JAM Press: Maple Grove, MN.

Linacre, J. M. (2010). Facets Rasch Measurement (Version 3.67.1) [Computer software]. Chicago: Winsteps.com.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of “scale analysis” and factor analysis. *Psychological Bulletin, 45*, 507-530.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246-276.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71.

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

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

McNamara, T. F. (2000). *Language Testing*. Oxford, UK: Oxford University Press.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*(4), 311-314.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, *14*(3), 283-298.

Meijer, R. R., & von Krimpen-Stoop, E. M. L. A. (2001). Person fit across subgroups: An achievement testing example. In A. Boomsma, M. A. J. van Duijin & T. A. B. Snijders (Eds.), *Essays on item response theory*, (pp. 377-390). New York: Springer-Verlag.

Messick, S. (1983). Assessment of children. In P. H. Mussen (Ed.), *Handbook of child psychology, Volume 1: History, theory and methods,* (pp. 477-526). New York, NY: John Wiley & Sons.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241-256.

Mislevy, R. J., Steinberg, L.S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15*(4), 363–389.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague: Mouton/Berlin: De Gruyter.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-367). New York: Springer.

Molenaar, I. W. (1982). Mokken scaling revisited. *Kwantitative Methoden, 3*(8), 145-164.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York: Springer.

Molenaar, I. W., & Sijtsma, K. (1984). Internal consistency and reliability in Mokken's nonparametric item response model. *Tijdschrift voor onderwijsresearch*, *9*, 257-268.

Molenaar, I. W., & Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve methoden*, *9*(28), 115-126.

Molenaar, I. W., & Sijtsma, K. (2000). MPS5 for Windows: A Program for Mokken Scale Analysis for Polytomous Items (Version 5.0) [Computer software]. Gronigen, The Netherlands: ProGAMMA.

Mullis, I. V. S., Martin, M. O., Foy, P., & Aora, A. (2012). *TIMSS 2011 International Results in Mathematics*. TIMSS & PIRLS International Study Center, Lynch School of Education: Chestnut Hill, MA, USA and International Association for the Evaluation of Educational Achievement (IEA): Amsterdam, the Netherlands.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *TIMSS 2011 International Results in Reading*. TIMSS & PIRLS International Study Center, Lynch School of Education: Chestnut Hill, MA, USA and International Association for the Evaluation of Educational Achievement (IEA): Amsterdam, the Netherlands.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*(4), 619.

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

Myford, C. M., Marr, D. B., & Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the Test of Written English*. (TOEFL Research Report No. 95-40). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs*. (TOEFL Technical Report, TR-15). Princeton, NJ: Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings. *Journal of Applied Measurement*, *3*(3), 300.

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

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 371-398.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371-389.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250.

OECD (2012). *PISA 2009 Technical Report, PISA*, OECD Publishing.

Penny, J. A., & Johnson, R. L. (2011). The accuracy of performance task scores after resolution of rater disagreement: A Monte Carlo study. *Assessing Writing*, *16*(4), 221-236.

Pula, J. J. & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Presskill, NJ: Hampton Press.

R Development Core Team (2013). R: A language and environment for statistical computing (Version 3.0.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

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. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (Vol. 4, pp. 321-333). Berkeley, CA: University of California Press.

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

Samejima, F. (1983). Some methods and approaches for estimating the operating characteristics of discrete item responses. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 159-182). Hillsdale, NJ: Erlbaum.

Samejima, F. (1997). The graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York: Springer.

Scarritt, J. R. (1996). Measuring political change: The quantity and effectiveness of electoral and party participation in the Zambian one-party state, 1973-91. *British Journal of Political Science*, *26*, 283-297.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1-30.

Scott, W. A. (1955) Reliability or content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19*, 321-325.

Seigel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw Hill.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks: Sage.

Shoukri, M. M. (2010). *Measures of interobserver agreement and reliability* (Vol. 39). Boca Raton: CRC Press, Taylor & Francis Group.

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

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 28* 65-94.

Sijtsma, K., Emons, W. H., Bouwmeester, S., Nyklíček, I., & Roorda, L. D. (2008). Nonparametric IRT analysis of quality-of-life scales and its application to the world health organization quality-of-life scale (WHOQOL-Bref). *Quality of Life Research*, *17*(2), 275-290.

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, *63*(2), 183-200.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149-157.

Sijtsma, K. A., & Meijer, R. R. (2007). Nonparametric item response theory and special topics. In C.R. Rao and S. Sinharay (Eds.), *Psychometrics, Handbook of statistics* (pp. 719-747),Volume 26. Amsterdam: Elsevier.

Sijtsma, K., Meijer, R. R., & van der Ark, L. A., (2011). Mokken scale analysis as time goes by: An update for scaling procedures. *Personality and Individual Differences, 50,* 31-37.

Sijtsma, K., & Molenaar, I. W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, *52*(1), 79-97.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory* (Vol. 5). Thousand Oaks: Sage.

Smith, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, *64*(4), 617-639.

Smith, R. M., Schumacker, R. E., & Bush, J. J. (2000). Examining replication effects in Rasch fit statistics. In M. Wilson & G. Engelhard, Jr. (Eds.). *Objective measurement: Theory into practice. Stamford, CT: Ablex Publishing Corp*, *5*, 303-317.

Stemler, S. E. and Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.) *Best practices in quantitative methods* (pp. 29-49). Los Angeles: Sage.

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*, 75-99.

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.

Thissen, D. & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements.* New York: Teachers College, Columbia University.

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

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 529-554.

Uebersax, J. S. (1992). Modeling approaches for the analysis of observer agreement. *Investigative Radiology*, *27*(9), 738.

Uebersax, J. (2002). *Statistical methods for rater agreement.* Retrieved February 4, 2013, from http://www.john-uebersax.com/stat/agree.htm#recs

US Department of Education (2010). *Race to the top assessment program executive summary.* Washington, D.C.: Author.

van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70,* 283-304.

van der Ark, L. A. (2010). Computation of the Molenaar Sijtsma statistic. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.) *Advances in data analysis, data handling and business intelligence* (pp. 775-784). Berlin: Springer-Verlag.

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48* (5), 1-27.

van der Ark, L. A. (2013). Mokken: Mokken scale analysis in R. *R package version 2.7.5* [Computer software]. URL http://www.jstatsoft.org/v20/i1 van der Ark, L.A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75,* 272-279.

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, *11*(2), 139-163.

van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis.* Los Angeles: Sage.

van Schuur, W. H. & Vis, J. C. P. M. (2000). What Dutch parliamentary journalists know about politics. *Acta Politica, 35*, 196-227.

Watson, R., Deary, I. J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine*, *38*(4), 575-580.

Weber E. H. (1846/1912). The sense of touch and common feeling. In B. Rand (Ed.). *The classical psychologists* (pp. 557–561). Boston: Houghton Mifflin

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

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 79*, 703-713.

Wiley, D. E., & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In M. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises, problems and challenges* (pp. 61-89). Mahwah, NJ: Lawrence Erlbaum Associates.

Wind, S. A. (2011). Rater-mediated domain response functions. *Rasch Measurement Transactions, 2011, 251:2, 1321-2*

Wind, S. A. & Engelhard (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement,13*(4), 321-335.

Wind, S. A. & Engelhard G., Jr. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*(4), 278-299.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67,* 189-205.

Wolcott, W. (with Legg, S. M.) (1998). *An overview of writing assessment: Theory, research, and practice*. Urbana, IL: National Council of Teachers of English.

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

Wolfe, E. W. (2009). Item and rater analysis of constructed response items via the multi-faceted Rasch model. *Journal of Applied Measurement*, *10*(3), 335-347.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31-37.

Wolfe, E. W., Moulder, B. C., & Myford, C. M. (2001). Detecting differential rater functioning over time (DRIFT) using a Rasch Multi-faceted rating scale model. *Journal of Applied Measurement*, *2*(3), 256-80.

Wolfe, E., Myford, C. M., Engelhard, G. Jr., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP English Literature and Composition examination using benchmark essays* (College Board Research Report No. 2007-2). New York: College Board.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement.* Chicago: MESA Press.

Wright, B. D., & Stone, M. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

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

Zhu, M. & Johnson, R. (2013, April). Robustness of inter-rater reliability estimators to rater leniency/severity effects in an absolute decision setting: A Monte Carlo study of performance ratings. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Zinn, F. D., Henderson, D. A., Nystuen, J. D., & Drake, W. D. (1992). A stochastic cumulative scaling method applied to measuring wealth in Indonesian villages. *Environment and Planning A*, *24*, 1155-1166.

Table 1. *Traditional Indices of Rating Quality*

Category	Definition
1. Rater agreement	• The degree to which raters assign equivalent scores to the same performance.
2. Rater errors and systematic biases	• 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.
3. Rater accuracy	• The degree to which raters assign scores equivalent to “true” or “known” scores.

Note. When indices of rating quality based on these three categories are applied, high levels of agreement, low levels of error and systematic bias, and high levels of accuracy are assumed to reflect high quality ratings.

Table 2. *Indices of Rater Agreement*

Types of Rater Agreement		**Questions**	**Statistics and Displays**
A. Categorical Agreement: Rater exchangeability	A1. Absolute agreement	Do raters assign the same scores to the same responses?	• Proportion of matching ratings assigned by rater pairs or across rater groups
	A2. Adjacent agreement	Do raters assign adjacent scores to the same responses?	• Proportion of adjacent ratings assigned by rater pairs or across rater groups
	A3. Chance-corrected agreement	Do raters assign matching ratings beyond what may be expected by chance alone?	• Kappa statistics
B. Ordinal Agreement: Relative consistency	B1. Correlation coefficients	How consistent is person ordering across a group of raters?	• Pearson product-moment correlation • Spearman's rho
	B2. Coefficient alpha	What is the internal consistency of ratings assigned by a group of raters?	• Cronbach's coefficient alpha coefficient
	B3. Intraclass correlations	What proportion of variation in ratings can be attributed to differences between raters?	• Various forms of ICC coefficients • Weighted kappa
	B4. Generalizability theory coefficients	To what degree do differences among raters contribute to measurement error?	• Variance components for raters • Standard error of measurement for raters • Generalizability and dependability coefficients

Note. This categorization of rater agreement indices is based on Stemler and Tsai (2008) and LeBrenton and Senter (2008).

Table 3. *Indices of Rater Error and Systematic Bias*

Types of Rater Error and Systematic Bias		Questions	Statistics and Displays
A. Distributional Errors: Discrepancies between assumed and observed distributions of “true” scores	A1. Leniency and severity	Do raters assign scores that are systematically higher or lower than warranted by student performance?	• Average ratings • Rater main effect in rater-by-student-by-domain ANOVA • Rating distribution (skewness)
	A2. Range restriction	Do raters discriminate among levels of achievement?	• Rater standard deviations across students within domains • Rating distribution (kurtosis) • Student main effect in rater-by-student-by-domain ANOVA • Rater standard deviations across students within domains
	A3. Central tendency	Do raters use the structure of the rating scale as intended?	• Rating distribution (kurtosis) • Student main effect in rater-by-student-by-domain ANOVA • Comparison of average ratings within a domain to the midpoint on a rating scale
B. Correlational errors: Distinguishing among distinct aspects of a performance	B1. Halo error	Do raters distinguish between distinct and independent rubric domains?	• Correlations among domain ratings • Standard deviations across domain ratings • Rater-by-student interaction in a rater-by-student-by-domain ANOVA • Factor analysis of a domain correlation matrix (percentage of variance accounted for by first principal component)
C. Systematic biases: Interactions between ratings and construct-irrelevant characteristics of assessments or students	C1. Interactions	Are ratings invariant over internal (e.g., prompts) and external (e.g., student subgroups) construct-irrelevant components?	• Interaction effects between raters and internal or external variables in ANOVA

Table 4. *Indices of Rater Accuracy*

Type of Rater Accuracy		**Questions**	**Statistics and Displays**
A. Categorical Accuracy: Match between operational and expert raters	A1. Distance accuracy	How close are observed ratings to expert ratings?	• Cronbach's D^2 index • Distance accuracy index
	A2. Accuracy components	What components of operational ratings match expert ratings?	• Elevation accuracy • Differential elevation accuracy • Stereotype accuracy • Differential accuracy
B. Ordinal accuracy: Similar ordering between operational and expert raters	B1. Correlational measures of accuracy	Does the rank ordering by observed ratings match that by expert ratings?	• Borman's (1977) distance accuracy

Table 5. *IRT Model Assumptions/Requirements*

Category	Question	Rasch Model Requirement	Mokken Model Assumptions	
			MH-R Model	**DM-R Model**
Dimensionality	How many latent variables are being measured?	***Unidimensionality:*** Responses reflect evidence of a single latent variable	***Unidimensionality:*** Responses reflect evidence of a single latent variable	***Unidimensionality:*** Responses reflect evidence of a single latent variable
Item Independence	What is the relationship among responses to individual items?	***Conditional independence:*** Responses to an item are not influenced by responses to any other item, after controlling for the latent variable	***Conditional independence:*** Responses to an item are not influenced by responses to any other item, after controlling for the latent variable	***Conditional independence:*** Responses to an item are not influenced by responses to any other item, after controlling for the latent variable
Functional Form	What is the mathematical relationship between person and item locations that describes the probability for observed responses?	***Parametric:*** The probability for observed responses follows a specific algebraic form such that item and person response functions do not intersect	***Nonparametric:*** the probability for observed responses does not need to conform to a specific shape, as long as the item response functions are monotonic	***Nonparametric:*** the probability for observed responses does not need to conform to a specific shape, as long as the item and person response functions are monotonic and do not intersect

Note. These categories are based on de Ayala (2009). Multiple assumptions can be included within each major category.

Table 6. *IRT Model Assumptions/Requirements for Rater Models*

Category	**Question**	**Rasch Model Requirement**	**Mokken Model Assumptions**	
			MH-R Model	**DM-R Model**
Dimensionality	How many latent variables are being measured?	***Unidimensionality:*** Ratings reflect evidence of a single latent variable	***Unidimensionality:*** Ratings reflect evidence of a single latent variable	***Unidimensionality:*** Ratings reflect evidence of a single latent variable
Rater Independence	What is the relationship among responses to individual items?	***Conditional independence:*** The rating assigned to a student is not influenced by ratings assigned by other raters	***Conditional independence:*** The rating assigned to a student is not influenced by ratings assigned by other raters	***Conditional independence:*** The rating assigned to a student is not influenced by ratings assigned by other raters
Functional Form	What is the mathematical relationship between student and rater locations that describes the probability for observed responses?	***Parametric:*** The probability for an observed rating follows a specific algebraic form such that rater and person response functions do not intersect	***Nonparametric:*** the probability for an observed rating does not need to conform to a specific shape, as long as the item response functions are monotonic	***Nonparametric:*** the probability for an observed rating does not need to conform to a specific shape, as long as the rater and person response functions are monotonic and do not intersect

Note. These categories are based on de Ayala (2009). Multiple assumptions can be included within each major category.

Table 7. *Rating Quality Indices based on the MFR Model for Ratings (Model I)*

Category	Rating Quality Indicator	Questions	Statistics and Displays
A. Rater Calibrations	• Rater leniency/severity	What is the location of each rater (severity/leniency)?	• Calibration and location of elements within facet • Variable map
	• Rater precision	How precisely has each rater been calibrated?	• Standard errors for raters
	• Rater separation	How spread-out are the individual rater severities? Can the raters be considered exchangeable?	• Reliability of separation statistic for raters • Chi square statistic for raters
B. Model-data Fit	• Model-data fit for raters	How consistently has each rater interpreted the domains and rating scale categories across students?	• Mean square error fit statistics (Outfit *MSE*)
C. Interactions	• Rater interactions	Are ratings invariant over internal construct-irrelevant components (e.g., prompts)?	• Interaction effects between rater and internal/external facets (bias analysis)

Note. This description of Rasch-based rating quality indices is based on Engelhard (2013).

Table 8. *Georgia Writing Results: Summary Statistics from MFR Model for Ratings (Model I)*

	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)	49,599.7*	869.5*	2,654.2*
(*df*)	(364)	(19)	(3)

* $p < .05$

Table 9. *Georgia Writing Results: Calibration of the Rater Facet from MFR Model for Ratings (Model I)*

Raters		Average Rating	Severity Measure (Logits)	SE	Infit MSE	Outfit MSE
Lenient	15	1.86	–0.56	0.05	1.00	1.10
↓	7	1.84	–0.51	0.05	1.21	1.22
	6	1.83	–0.46	0.05	0.97	0.98
	10	1.80	–0.35	0.04	0.82	0.81
	5	1.78	–0.28	0.05	1.17	1.16
	12	1.76	–0.18	0.05	0.87	0.89
	11	1.73	–0.09	0.05	0.82	0.80
	21	1.73	–0.06	0.05	1.26	1.21
	2	1.72	–0.03	0.05	0.80	0.78
	3	1.72	–0.02	0.05	0.97	1.05
	19	1.70	0.05	0.05	1.24	1.21
	4	1.69	0.08	0.05	0.81	0.79
	17	1.69	0.10	0.05	0.81	0.80
	20	1.68	0.14	0.05	1.00	1.01
	13	1.67	0.16	0.05	1.00	0.99
	14	1.67	0.16	0.05	1.02	1.06
	8	1.65	0.25	0.05	0.93	0.91
	16	1.61	0.40	0.05	0.97	1.06
	18	1.58	0.49	0.05	1.11	1.27
	9	1.52	0.73	0.05	0.76	0.73
Severe						
	Mean	1.71	0.00	0.05	0.98	0.99
	SD	0.09	0.33	0.00	0.16	0.17

Note. Raters are ordered by Severity Measure from low (lenient) to high (severe).

Table 10. *Rating Quality Indices based on the MFR Model for Rater Accuracy (Model II)*

Category	Rating Quality Indicator	Questions	Statistics and Displays
A. Rater Accuracy Calibrations	• Rater leniency/severity accuracy	What is the accuracy location of each rater?	• Variable map • Calibration and location of elements within facet
	• Rater accuracy precision	How precisely has each rater been calibrated in terms of accuracy?	• Standard errors for raters
	• Rater accuracy separation	How spread out are the individual raters in terms of accuracy?	• Reliability of separation statistic for raters
		Can the raters be considered exchangeable in terms of accuracy?	• Chi square statistic for raters
B. Model-data Fit	• Model-data fit for rater accuracy	How consistently does each rater demonstrate accuracy across the domains, rating scale categories, and students?	• Mean square error fit statistics (Infit and Outfit *MSE*)
C. Interactions	• Rater accuracy interactions	Is rater accuracy invariant over internal construct-irrelevant components (e.g., prompts)?	• Interaction effects between rater and internal facets (bias analysis)

Note. This description of Rasch-based rating quality indices is based on Engelhard (2013).

Table 11. *Georgia Writing Results: Summary Statistics from MFR Model for Rater Accuracy (Model II)*

	Rater Accuracy (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)	622.6*	1,348.6*	18.6*
(*df*)	(19)	(364)	(3)

* $p < .05$

Note. The Benchmark Papers (δ) and Domains (α) facets represent difficulty for accurate ratings on individual papers and domains, respectively.

Table 12. *Georgia Writing Results: Calibration of the Rater Facet from MFR Model for Rater Accuracy (Model II)*

Raters		**Average Accuracy Score**	**Accuracy Measure (Logits)**	**SE**	**Infit MSE**	**Outfit MSE**
(Low accuracy)	19	0.60	-0.37	0.06	1.00	0.98
↓	7	0.60	-0.35	0.06	0.99	0.98
	21	0.61	-0.32	0.06	1.01	1.01
	18	0.61	-0.29	0.06	1.00	1.01
	5	0.62	-0.27	0.06	1.00	0.98
	15	0.64	-0.17	0.06	1.02	1.02
	3	0.65	-0.14	0.06	1.03	1.05
	16	0.65	-0.10	0.06	1.00	1.00
	9	0.66	-0.09	0.06	1.00	0.98
	13	0.66	-0.06	0.06	0.99	0.96
	8	0.67	-0.04	0.06	0.99	1.02
	20	0.67	-0.04	0.06	1.00	1.03
	14	0.68	0.00	0.06	1.00	1.00
	17	0.69	0.05	0.06	1.00	0.99
	11	0.69	0.07	0.06	0.96	0.93
	12	0.69	0.07	0.06	0.99	0.99
	6	0.73	0.26	0.06	1.02	1.04
	4	0.75	0.38	0.06	1.00	1.00
	2	0.78	0.56	0.06	1.00	0.98
(High accuracy)	10	0.82	0.85	0.05	1.01	1.06
	Mean	0.67	-0.04	0.06	1.00	1.00
	SD	0.06	0.31	0.00	0.01	0.03

Note. Raters are ordered by Accuracy Measure (logit scale) from low (inaccurate) to high (accurate).

Table 13. *Rating Quality indices based on Mokken Scaling*

Category	Rating Quality Indicator	Question(s)	Statistics and Plots
A. Scalability	Rater scalability	Can individual raters distinguish among students across achievement levels?	• Individual rater scalability coefficients • Rater pair scalability coefficients • Group rater scalability coefficients
B. Monotone Homogeneity	Rater monotonicity	Does the group of raters share the same relative ordering of students across achievement levels?	• Monotonicity plots and statistics for overall raters • Monotonicity plots and statistics within rating scale categories
C. Double monotonicity	Rater double monotonicity	• Is the ordering of students in terms of the latent variable invariant across raters? Is the ordering of raters in terms of the latent variable invariant across students?	• Pairwise rater restscore plots for overall raters • Pairwise cumulative category probability plots
D. Invariant ordering	Manifest invariant rater ordering (Ligtvoet et al., 2010, 2011)	• Is the relative ordering of rating scale categories for raters consistent across achievement levels?	• Manifest invariant ordering statistics and pairwise plots
E. Reliability	• Rater reliability	• What proportion of observed variance is attributable to true score variance?	• Molenaar and Sijtsma (1985, 1987) reliability statistic

Table 14. *Georgia Writing Results: Rating Quality indices based on Mokken Scaling*

	NIRT Rating Quality Indices					
	A. Scalability		**B. Monotonicity**	**C. Double Monotonicity**		**D. Invariant Ordering**
Assessment Opportunity	Rater scalability coefficient H_i (SE)	Number of negative rater pair scalability coefficients	Number of violations (number significant)	Number of violations via restscore method (number significant)	Number of violations via P matrix method (number significant)	Number of violations via MIRO method (number significant)
Rater 2	0.77 (0.02)	0	0 (0)	0 (0)	0 (0)	2 (0)
Rater 3	0.76 (0.02)	0	0 (0)	5 (0)	0 (0)	1 (0)
Rater 4	0.78 (0.02)	0	0 (0)	6 (1)	0 (0)	1(0)
Rater 5	0.77 (0.02)	0	0 (0)	5 (2)	1 (1)	1 (0)
Rater 6	0.76 (0.02)	0	0 (0)	2 (0)	0 (0)	3 (0)
Rater 7	0.74 (0.02)	0	0 (0)	6 (0)	1 (1)	1(0)
Rater 8	0.78 (0.02)	0	0 (0)	3 (0)	1 (1)	2 (1)
Rater 9	0.82 (0.02)	0	0 (0)	0 (0)	0 (0)	0 (0)
Rater 10	0.78 (0.02)	0	0 (0)	2 (0)	0 (0)	1 (0)
Rater 11	0.78 (0.02)	0	0 (0)	2 (0)	0 (0)	3 (0)
Rater 12	0.78 (0.02)	0	0 (0)	0 (0)	0 (0)	3 (0)
Rater 13	0.78 (0.02)	0	0 (0)	3 (0)	2 (2)	1 (0)
Rater 14	0.76 (0.02)	0	0 (0)	5 (0)	1 (1)	3 (0)
Rater 15	0.77 (0.02)	0	0 (0)	2 (0)	0 (0)	0 (0)
Rater 16	0.78 (0.02)	0	0 (0)	3 (1)	1 (1)	6 (3)
Rater 17	0.80 (0.02)	0	0 (0)	4 (1)	0 (0)	7 (2)
Rater 18	0.75 (0.02)	0	0 (0)	8 (0)	3 (3)	2 (0)
Rater 19	0.76 (0.02)	0	0 (0)	6 (0)	1 (1)	0 (0)
Rater 20	0.78 (0.02)	0	0 (0)	5 (2)	2 (2)	3 (1)
Rater 21	0.74 (0.02)	0	0 (0)	9 (1)	1 (1)	2 (1)

Table 15. *Rating Scale Guidelines*

Guidelines		Rasch Indices (Parametric)	Mokken Indices (Nonparametric)
1. Directional orientation with the latent variable:	Increasing amounts of a latent variable (θ) correspond to increasing categories on a rating scale.	A. Monotonically increasing expected score ogive	B. Average ratings increase monotonically across rest scores C. Category response functions increase monotonically across rest scores
2. Category precision:	Rating scale categories are distinct.	A. Normal/uniform distribution of ratings across categories B. |Difference| between category coefficient locations (logit scale) between ~1.4 and 5.0 logits C. Multimodal category probability functions D. Conditional probability curves are distinct and evenly spaced along the logit scale E. Smooth item information functions F. Smooth category information functions	G. Category response functions do not overlap *within* items H. Category response functions do not overlap *across* items
3. Model-data fit:	Rating scale categories meet the expectations of models with useful measurement properties.	A. Close match between observed and expected score ogives B. Outfit *MSE* statistics for categories are near their expected value (~1.00)	C. Item scalability coefficients (H_i) suggest scalable items (≥ ~ 0.3) D. Category response functions do not intersect *across* items E. Manifest invariant item ordering is observed (Ligtvoet et al., 2010)

Note. These guidelines are adapted from Linacre (1999, 2004). Each indicator of rating scale effectiveness is examined for the total group of students (N = 8,620), and within the gender and race/ethnicity subgroups.

Table 16. *Summary Statistics for Rasch Partial-Credit Model (Overall sample; N=8,620)*

		Students	ECR items
Measures			
	M	−0.58	0.00
	SD	2.20	0.17
	N	8,620	4
Infit *MSE*			
	M	0.93	0.95
	SD	0.73	0.04
Outfit *MSE*			
	M	0.95	0.96
	SD	0.75	0.04
Std. Infit *MSE*			
	M	−0.20	−4.10
	SD	1.30	3.40
Std. Outfit *MSE*			
	M	−0.20	−3.80
	SD	1.30	3.40
Reliability of Separation		0.89	0.99
χ^2 Statistic		73,701.1*	690.80*
df		8,619	3

* $p < 0.05$

Table 17. *Guideline 2: Indices of Category Precision*

	Rasch (Parametric)															
Indicator	A. Normal/uniform distribution of ratings across categories				B. |Difference| between category coefficient locations (logit scale) between ~1.40 and 5.0 logits				C. Multimodal category response functions				D. Conditional probability curves are evenly spaced along the logit scale			
	ECR1	ECR2	ECR3	ECR4	ECR1	ECR2	ECR3	ECR4	ECR1	ECR2	ECR3	ECR4	ECR1	ECR2	ECR3	ECR4
Total	1	--	1	1	--	--	--	--	--	--	--	--	3	--	3	3
Female	1	--	1	1	--	2	--	--	--	--	--	--	3	--	3	3
Male	1	--	1	1	--	--	--	--	--	--	--	--	3	--	3	3
AK Native	1	--	1	1	--	--	--	--	--	--	--	--	3	--	3	3
White	1	--	1	1	--	--	--	--	--	--	--	--	3	--	3	3

	Rasch (Parametric)								Mokken (Nonparametric)							
Indicator	E. Smooth item information functions				F. Smooth category information functions				G. Category response functions do not overlap *within* items				H. Category response functions do not overlap *across* items			
	ECR1	ECR2	ECR3	ECR4	ECR1	ECR2	ECR3	ECR4	ECR1	ECR2	ECR3	ECR4	ECR1	ECR2	ECR3	ECR4
Total	4	--	5	6	9	--	9	10	--	--	--	--	13	14	15	16
Female	6	--	9	7	9	--	9	10	--	11	--	--	13	14	15	16
Male	6	--	7	6	9	--	9	10	--	12	--	--	13	14	15	16
AK Native	6	8	7	7	9	10	9	10	--	--	--	--	13	14	15	16
White	6	--	6	6	9	--	9	10	--	--	--	--	13	14	15	16

Notes. The "- -" entry indicates no evidence for violation of the guideline. Observed violations are as follows:1) Distribution is left-skewed; 2) Difference between the first two rating scale categories ≥ |5.00| logits; 3) "Gap" along the logit scale between the first two conditional probability curves; 4) Reduced information between -5 and -1 logits; 5) Reduced information between -5 and -2 logits; 6) Reduced information between -3 and -1 logits; 7) Reduced information between -3 and 0 logits; 8) Reduced information between -6 and -3 logits; 9) Category 2 is bimodal; 10) Category 2 is flat; 11) Categories 1 and 2 overlap for students with R ≥ 14; 12) Categories 1 and 2 overlap for students with R ≥ 15; 13) Cumulative category probabilities overlap with ECR3 and ECR4; 14) Cumulative category probabilities overlap with ECR1, ECR3, & ECR4 in highest rating scale category; 15) Cumulative category probabilities overlap with ECR1 and ECR 4; 16) Cumulative category probabilities overlap with ECR1 and ECR3. Additional details about these violations are provided in the text.

Table 18. *Indices of Model-Data Fit*

	Rasch (Parametric)								**Mokken (Nonparametric)**											
Indi-cator	A. Close match between observed and expected score ogives				B. Outfit *MSE* statistics for categories are near their expected value (~1.00)				C. Item scalability coefficients (H_i) suggest scalable items				D. Category response functions do not intersect *across* items				E. Manifest invariant item ordering			
	ECR 1	ECR 2	ECR 3	ECR 4	ECR 1	ECR 2	ECR 3	ECR 4	ECR 1	ECR 2	ECR 3	ECR 4	ECR 1	ECR 2	ECR 3	ECR 4	ECR 1	ECR 2	ECR 3	ECR 4
Total	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
Female	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	4	4
Male	1	--	--	--	--	--	--	--	--	--	--	--	--	--	3	3	4	--	4	4
AK Native	--	--	--	--	--	--	--	--	--	--	--	--	3	--	3	--	4	--	4	4
White	--	--	2	--	--	--	--	--	--	--	--	--	3	--	3	--	--	--	--	--

Notes. The "- -" entry indicates no evidence for violation of the guideline. Observed violations are as follows: 1) Unexpected ratings near 5 logits; 2) Unexpected ratings near 6 logits; 3) Violation of nonintersecting cumulative category probabilities is statistically significant; 4) Violation of Manifest Invariant Item Ordering is statistically significant. Additional details about these violations are provided in the text.

Figure 1. *Framework for Evaluating the Quality of Rater-Mediated Assessments*

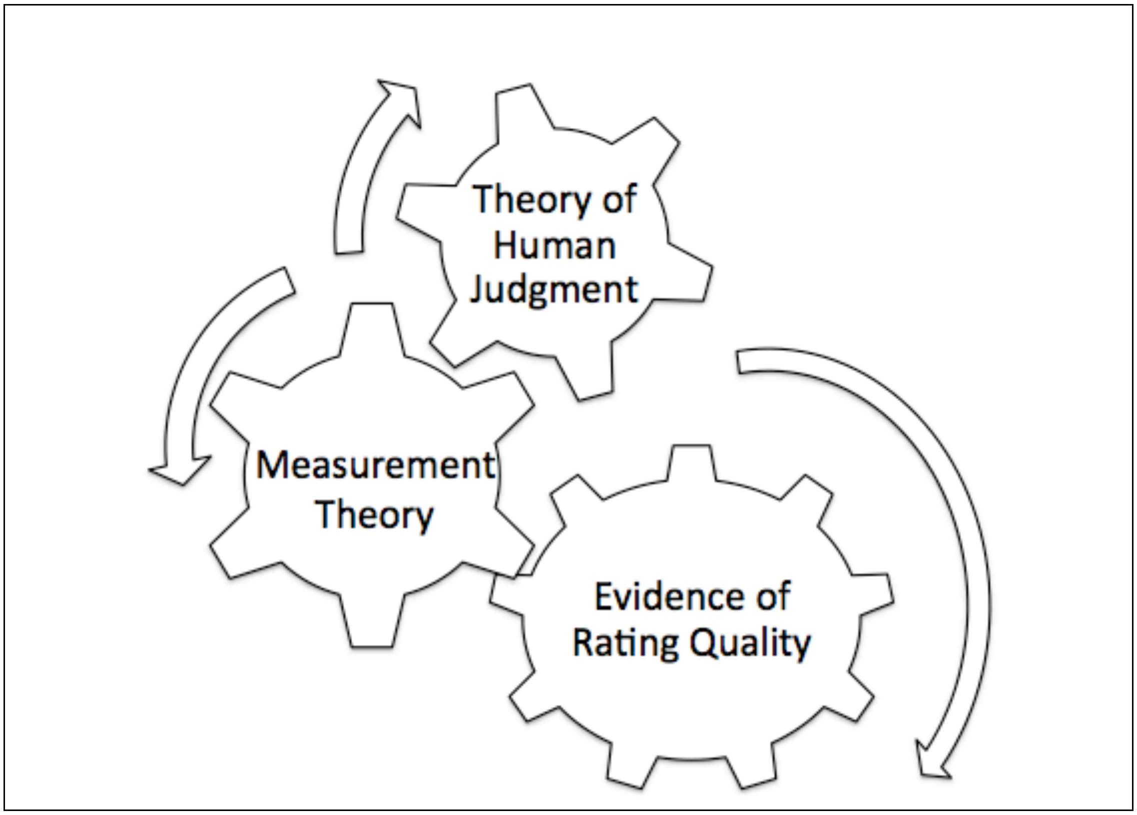
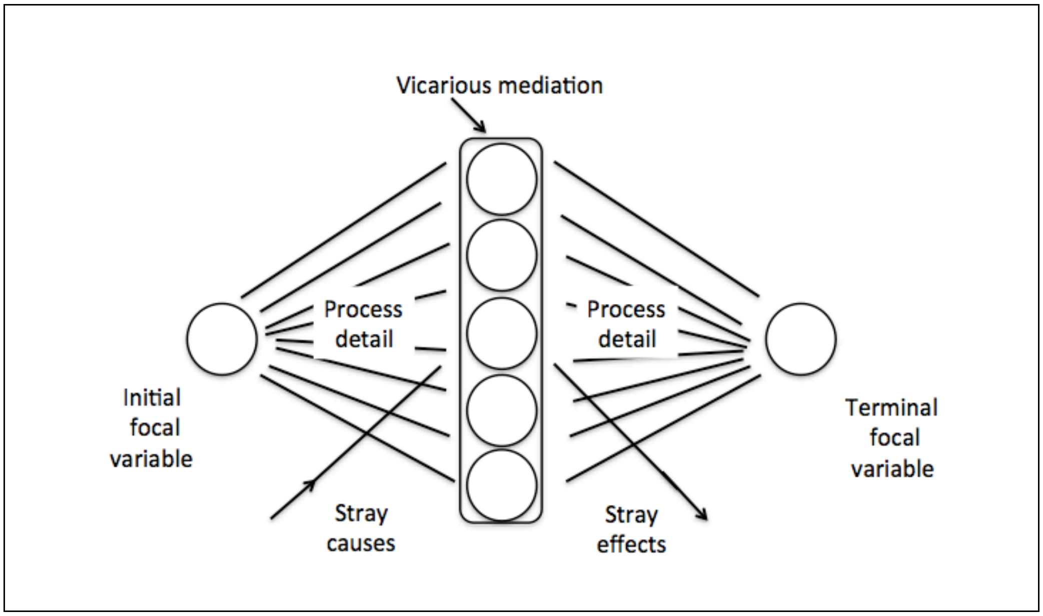
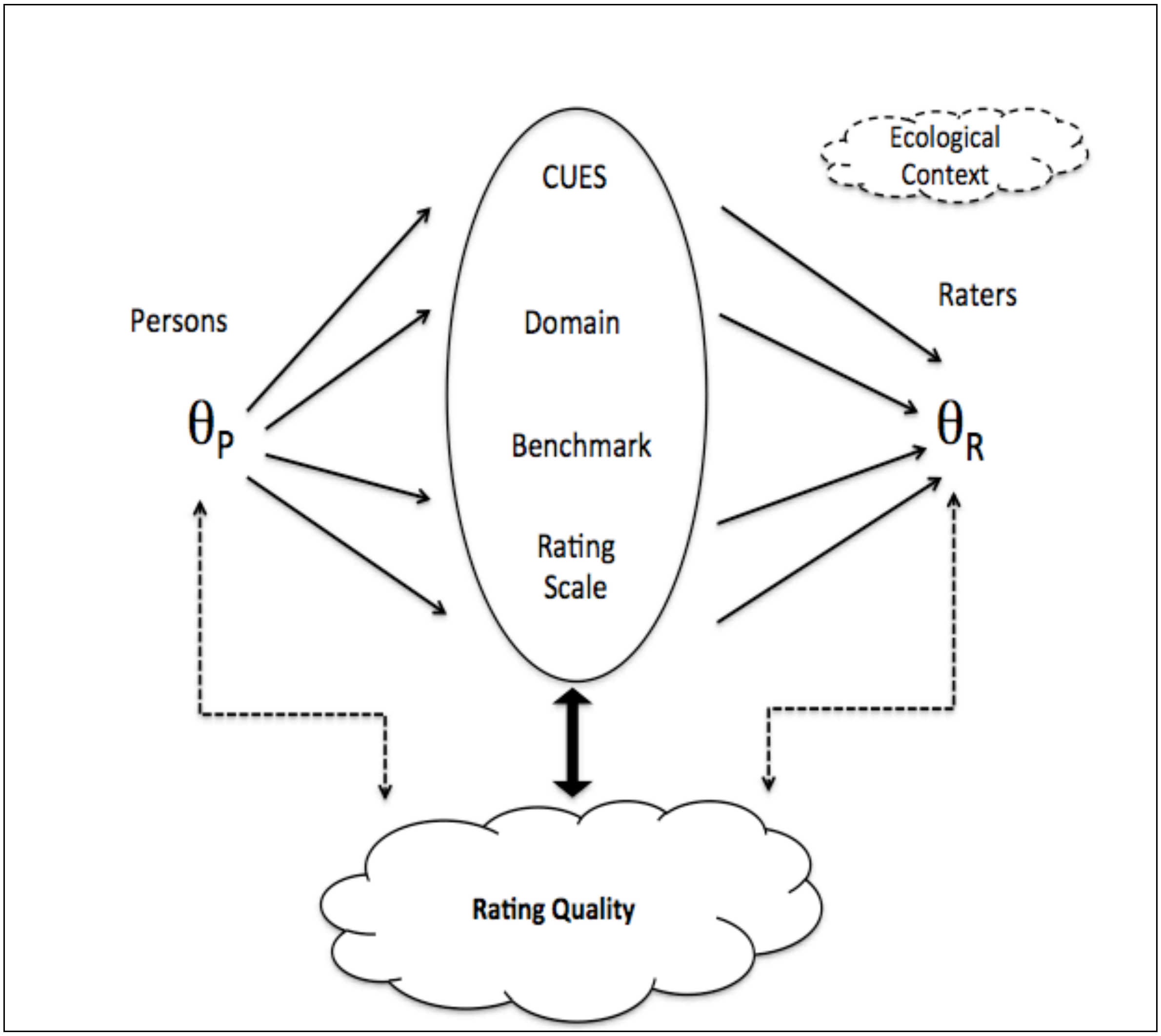


Figure 2. *Brunswik's Lens Model for Probabilistic Functionalism*



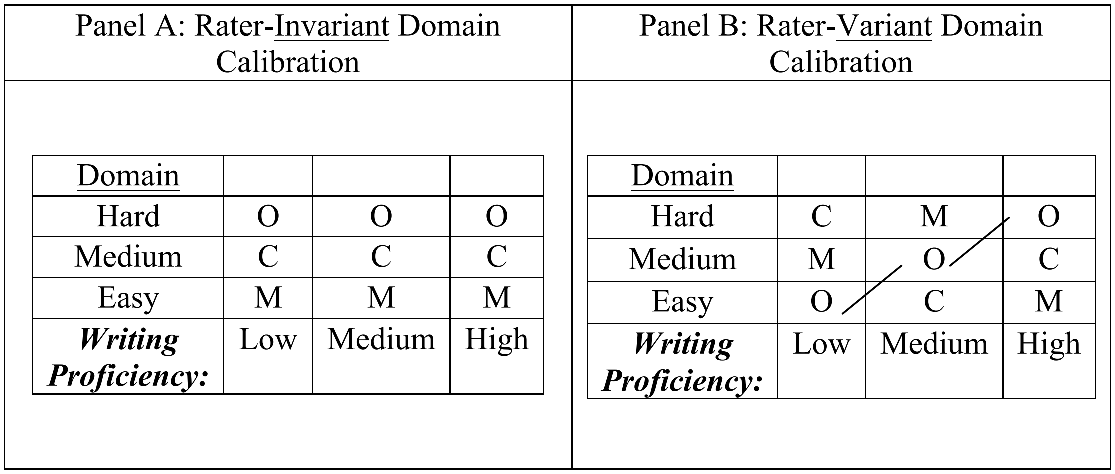
Note. Adapted from Brunswik (1952).

Figure 3. *Lens Model for Rater-Mediated Assessment*



Note. This lens model is an adaptation of Engelhard's (2013) lens model for rater-mediated assessments.

Figure 4. *Rater-Invariant Domain Calibrations*



Panel A: Rater-Invariant Domain Calibration

Domain			
Hard	O	O	O
Medium	C	C	C
Easy	M	M	M
Writing Proficiency:	Low	Medium	High

Panel B: Rater-Variant Domain Calibration

Domain			
Hard	C	M	O
Medium	M	O	C
Easy	O	C	M
Writing Proficiency:	Low	Medium	High

Note. The three domains are Mechanics (M), Organization (O), and Content (C).

Figure 5. *Rater-Variant Rating Scale Calibration*

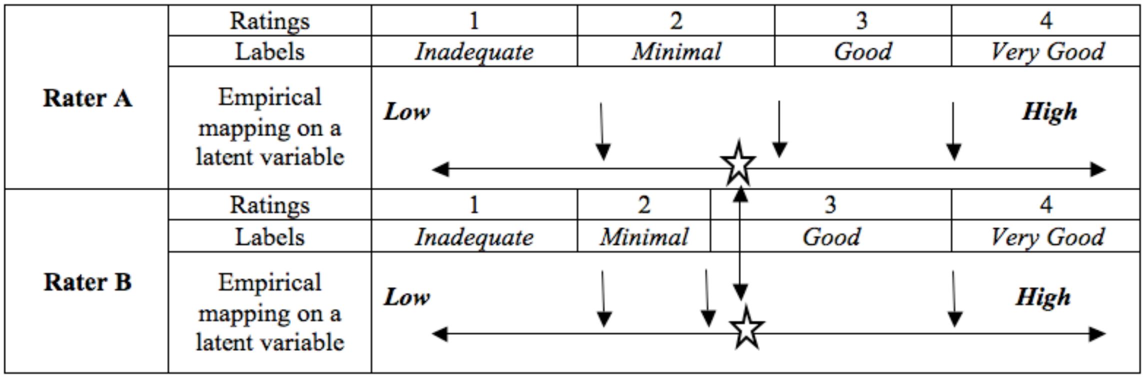


Figure 6. *Variable Map for Writing Assessment*

Person Locations	Logit Scale	CUES			Rater	Judged Person Locations
		Domains	Benchmarks	Rating Categories		
High	**4.00**	*Conventions*	High Proficiency	*High*		*High*
	3.00					
θ_A →	**2.00**					← Essay A
	1.00	*Organization*	Medium Proficiency	*Medium*		
θ_B →	**.00**				Rater λ	← Essay B
	-1.00	*Ideas*				
θ_C →	**-2.00**		Low Proficiency	*Low*		← Essay C
	-3.00	*Style*				
Low	**-4.00**					*Low*

Figure 7. *Rasch Operating Characteristic Functions (OCFs)/Item Response Functions (IRFs)*

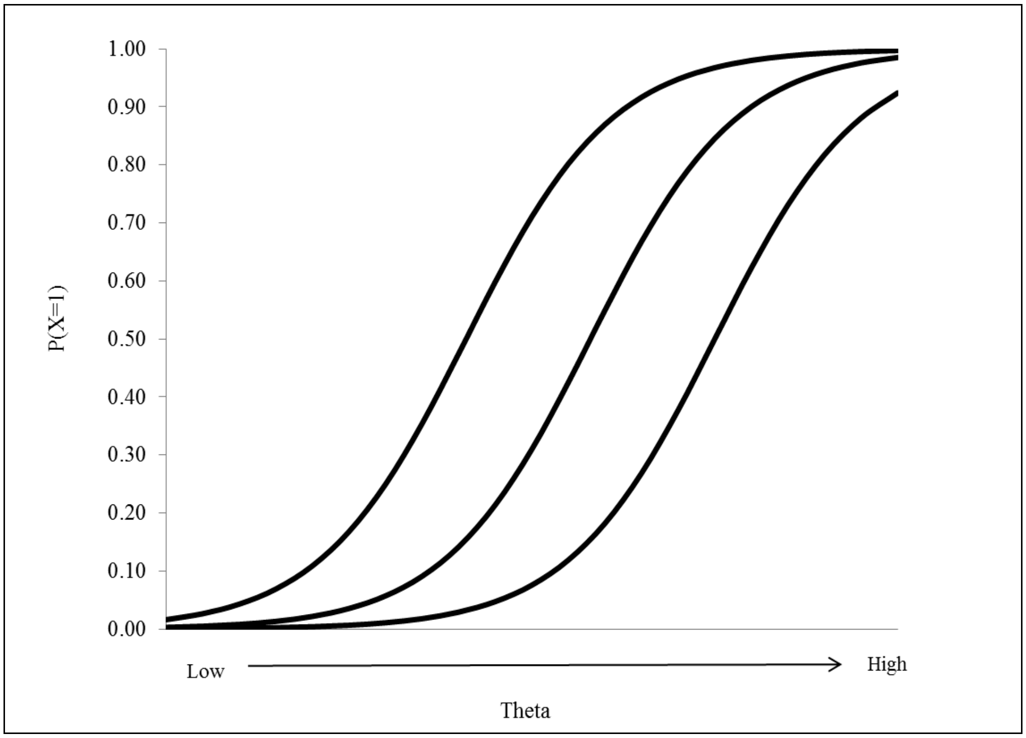
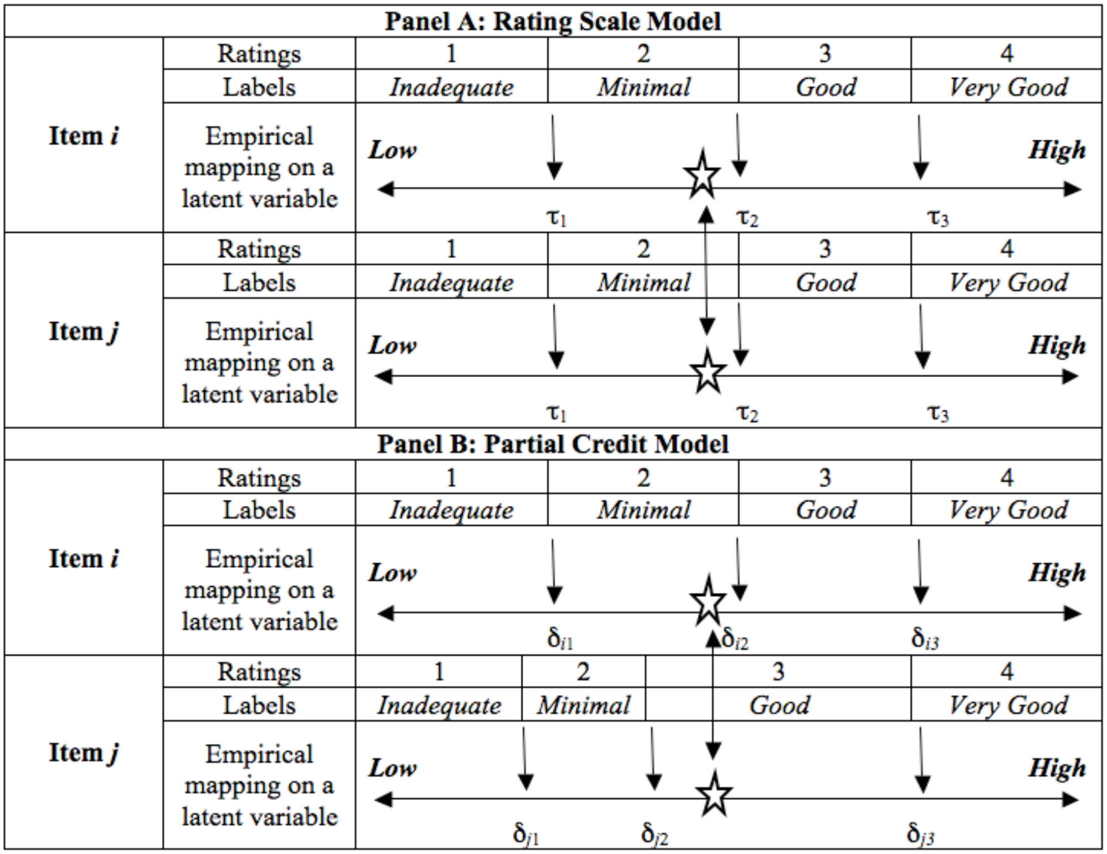
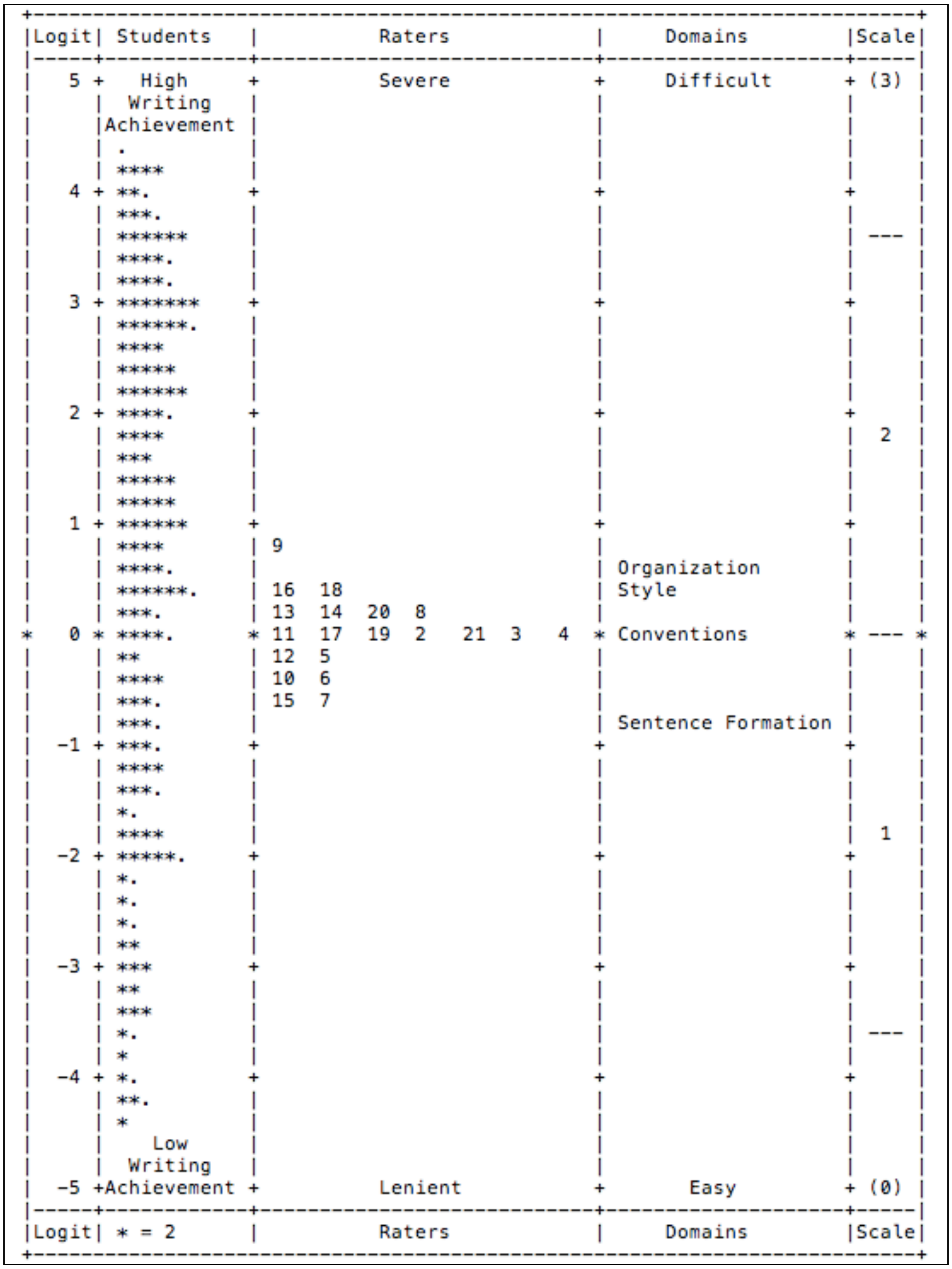


Figure 8. *Calibration of Category Coefficients under the Rasch Rating Scale and Partial-Credit Models*



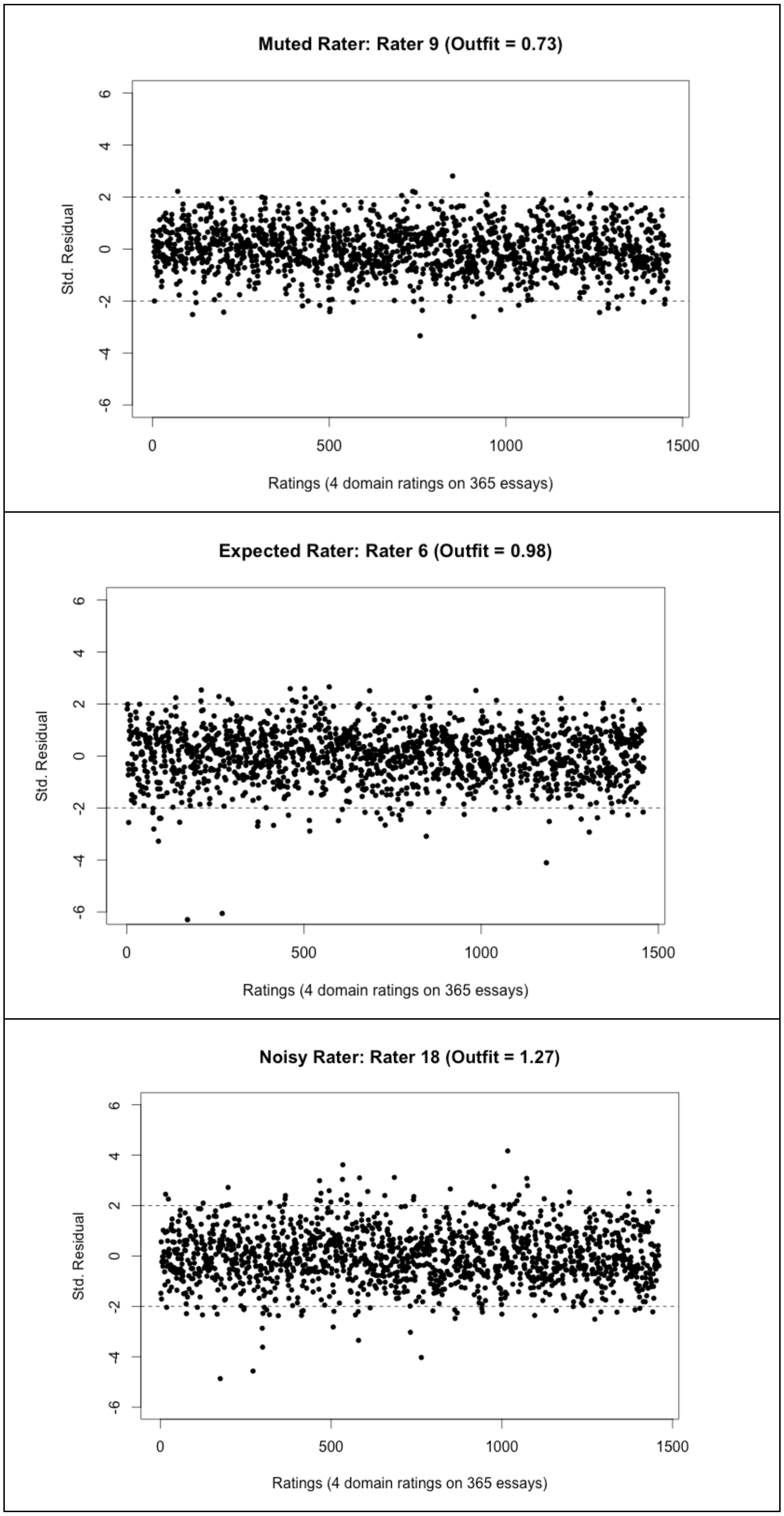
Note. The stars represent the latent-variable location of a student.

Figure 9. *Georgia Writing Results: Variable Map for MFR Model for Ratings (Model I)*



```
+--------------------------------------------------------------------------+
|Logit| Students   |         Raters             |     Domains        |Scale|
|-----+------------+----------------------------+--------------------+-----|
|   5 +   High     +          Severe            +     Difficult      + (3) |
|     |  Writing   |                            |                    |     |
|     |Achievement |                            |                    |     |
|     | .          |                            |                    |     |
|     | ****       |                            |                    |     |
|   4 + **.        +                            +                    +     |
|     | ***.       |                            |                    |     |
|     | ******     |                            |                    | --- |
|     | ****.      |                            |                    |     |
|     | ****.      |                            |                    |     |
|   3 + *******    +                            +                    +     |
|     | ******.    |                            |                    |     |
|     | ****       |                            |                    |     |
|     | *****      |                            |                    |     |
|     | ******     |                            |                    |     |
|   2 + ****.      +                            +                    +     |
|     | ****       |                            |                    |  2  |
|     | ***        |                            |                    |     |
|     | *****      |                            |                    |     |
|     | *****      |                            |                    |     |
|   1 + ******     +                            +                    +     |
|     | ****       | 9                          |                    |     |
|     | ****.      |                            | Organization       |     |
|     | ******.    | 16  18                     | Style              |     |
|     | ***.       | 13  14  20  8              |                    |     |
*   0 * ****.      * 11  17  19  2   21  3   4  * Conventions        * --- *
|     | **         | 12  5                      |                    |     |
|     | ****       | 10  6                      |                    |     |
|     | ***.       | 15  7                      |                    |     |
|     | ***.       |                            | Sentence Formation |     |
|  -1 + ***.       +                            +                    +     |
|     | ****       |                            |                    |     |
|     | ***.       |                            |                    |     |
|     | *.         |                            |                    |     |
|     | ****       |                            |                    |  1  |
|  -2 + *****.     +                            +                    +     |
|     | *.         |                            |                    |     |
|     | *.         |                            |                    |     |
|     | *.         |                            |                    |     |
|     | **         |                            |                    |     |
|  -3 + ***        +                            +                    +     |
|     | **         |                            |                    |     |
|     | ***        |                            |                    |     |
|     | *.         |                            |                    | --- |
|     | *          |                            |                    |     |
|  -4 + *.         +                            +                    +     |
|     | **.        |                            |                    |     |
|     | *          |                            |                    |     |
|     |    Low     |                            |                    |     |
|     |  Writing   |                            |                    |     |
|  -5 +Achievement +          Lenient           +      Easy          + (0) |
|-----+------------+----------------------------+--------------------+-----|
|Logit| * = 2      |         Raters             |     Domains        |Scale|
+--------------------------------------------------------------------------+
```

Figure 10. *Georgia Writing Results: Standardized Residual Plots for Observed Ratings for Three Raters with Different Levels of Model-data Fit based on MFR Model for Ratings (Model I)*

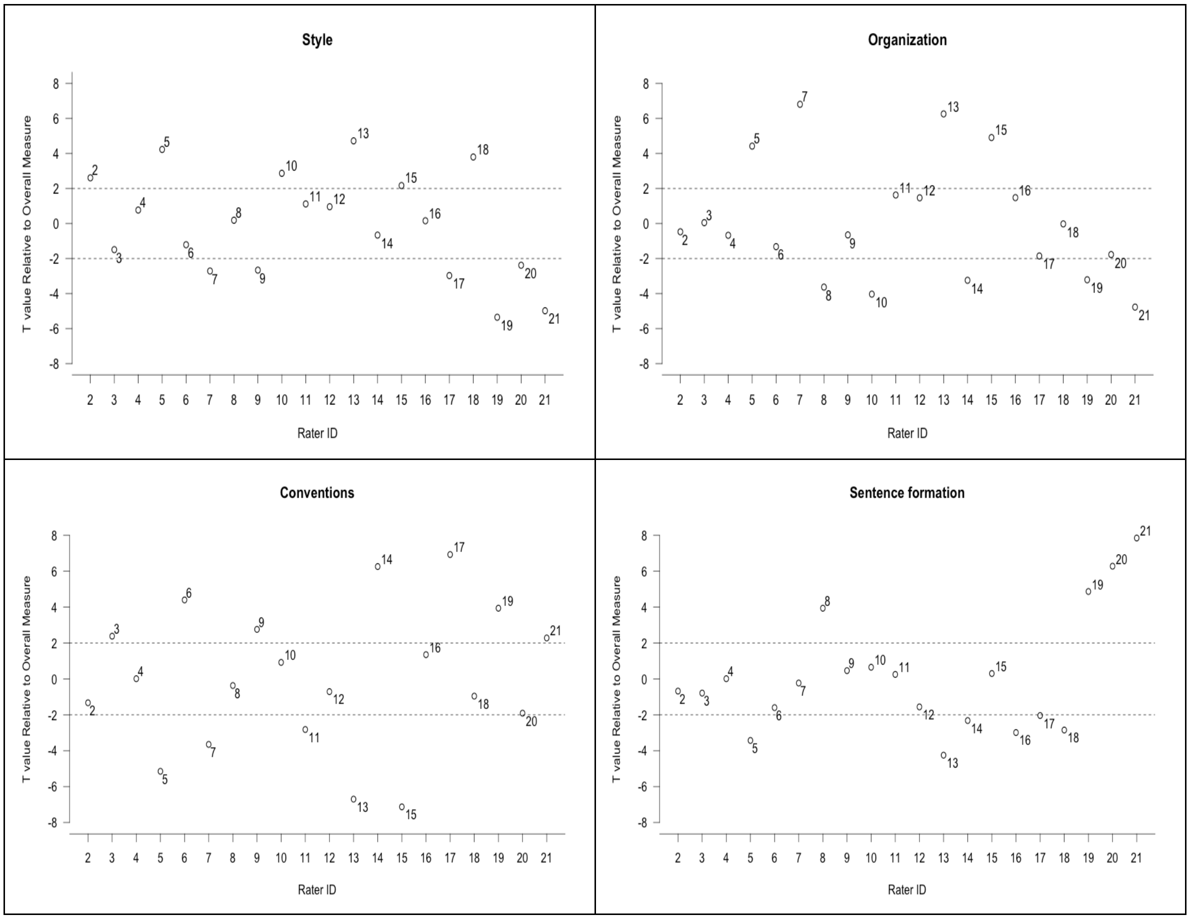


Muted Rater: Rater 9 (Outfit = 0.73)

Expected Rater: Rater 6 (Outfit = 0.98)

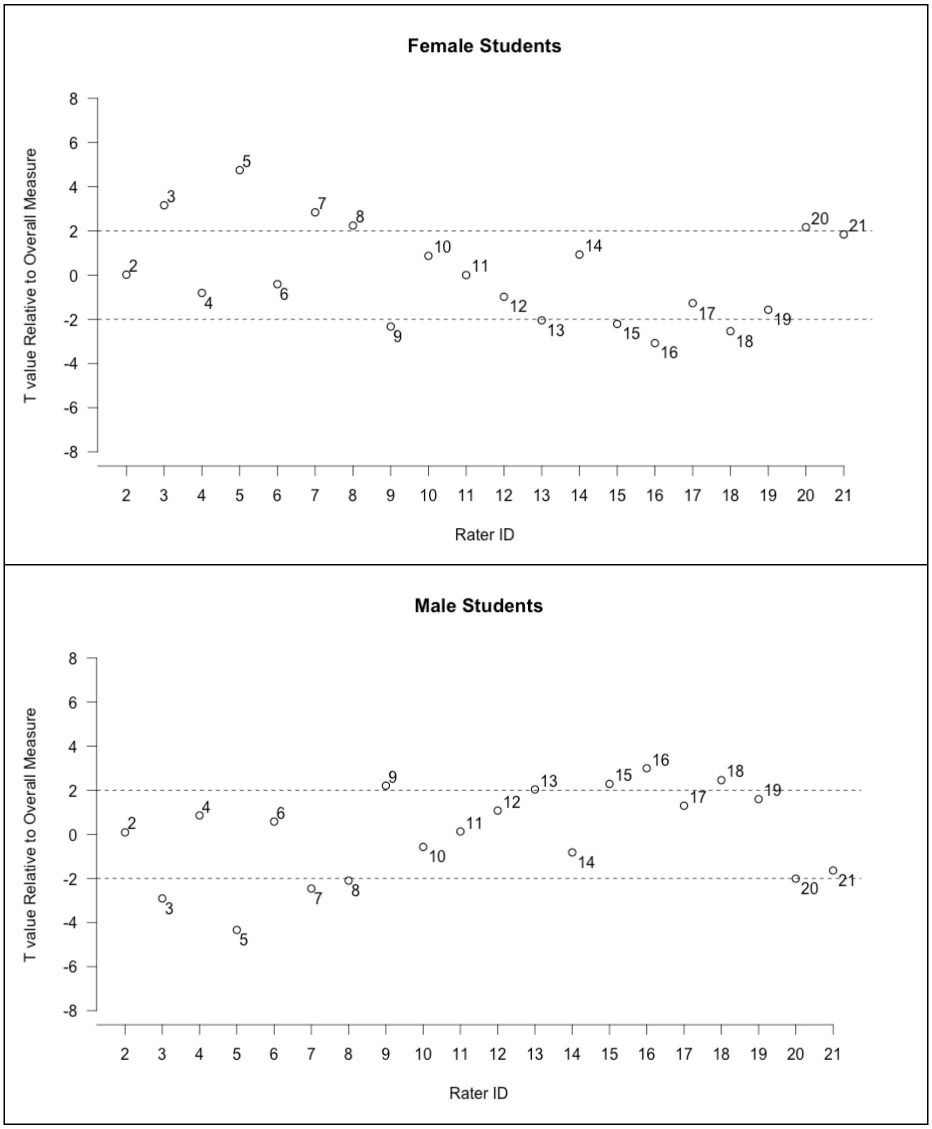
Noisy Rater: Rater 18 (Outfit = 1.27)

Figure 11. *Georgia Writing Results: Interactions between Rater Severity and Domain Difficulty based on MFR Model for Ratings (Model I)*



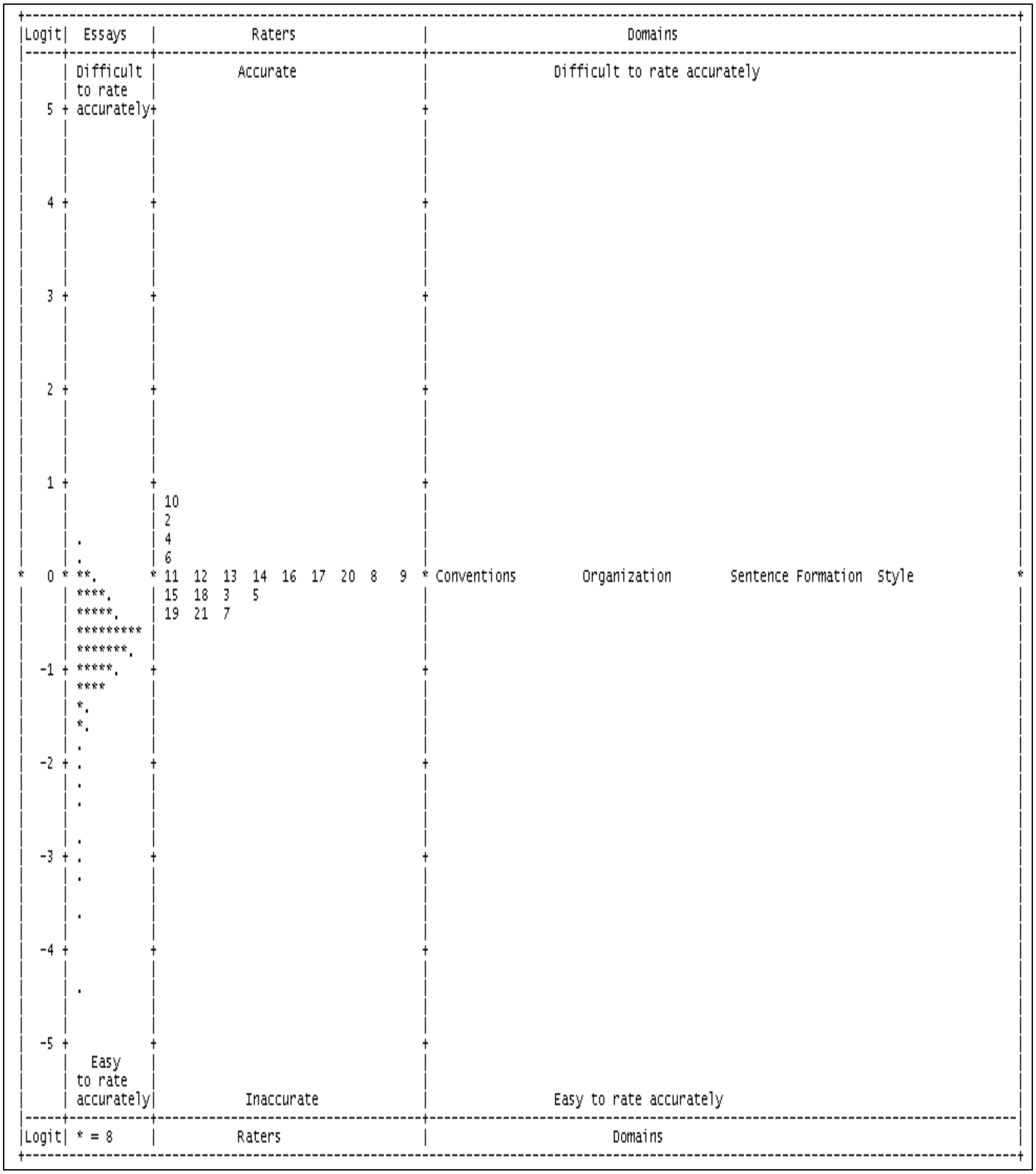
Note. Values of the *t*-statistic shown here are tests of the hypothesis that there is no interaction between rater severity and domain difficulty. Values higher than +2.00 suggest that the rater assigned higher ratings (i.e., was more lenient) than expected on a domain, based on its overall judged difficulty measure across the raters. Test statistic values lower than -2.00 suggest that the rater assigned lower ratings (i.e., was more severe) than expected on a domain.

Figure 12. *Georgia Writing Results: Interactions between Rater Severity and Student Gender based on MFR Model for Ratings (Model I)*



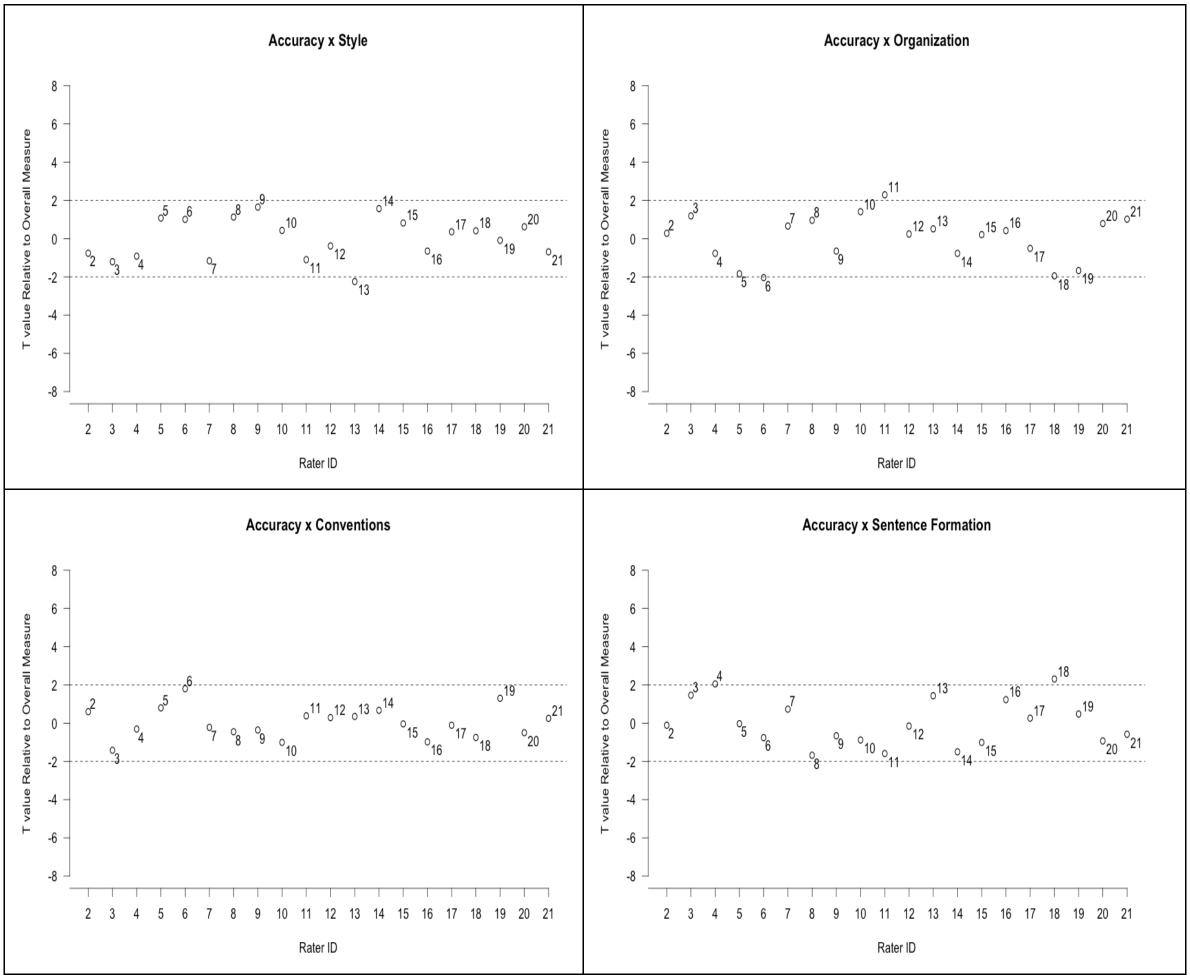
Note. Values of the *t*-statistic shown here are tests of the hypothesis that there is no interaction between rater severity and student gender. Values higher than +2.00 suggest that the rater assigned higher ratings (i.e., was more lenient) than expected for a subgroup of students, based on the average measure for the subgroup across the raters. Test statistic values lower than -2.00 suggest that the rater assigned lower ratings (i.e., was more severe) than expected for a subgroup of students.

Figure 13. *Georgia Writing Results: Variable map based on MFR Model for Rater Accuracy (Model II)*



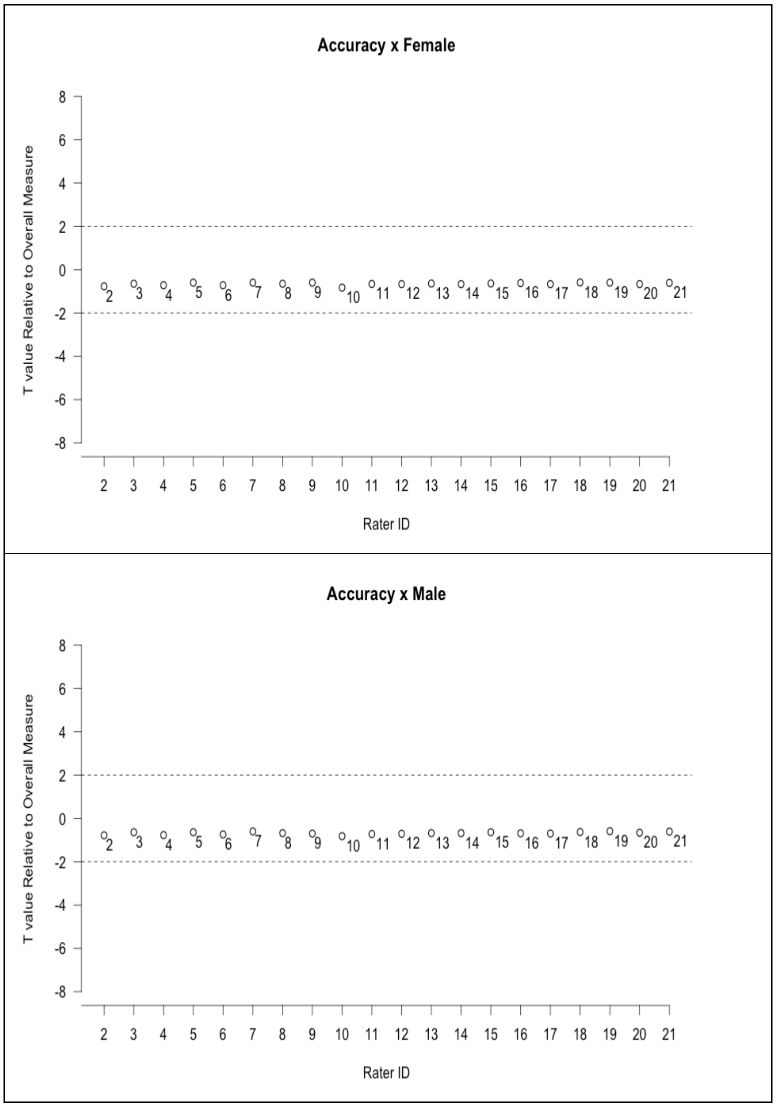
```
+-------------------------------------------------------------------------------------------------------------------------------------+
|Logit|  Essays   |             Raters                 |                           Domains                                              |
|-----+-----------+------------------------------------+--------------------------------------------------------------------------------|
|     | Difficult |           Accurate                 |                 Difficult to rate accurately                                   |
|     | to rate   |                                    |                                                                                |
|   5 + accurately+                                    +                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|   4 +           +                                    +                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|   3 +           +                                    +                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|   2 +           +                                    +                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|   1 +           +                                    +                                                                                |
|     |           | 10                                 |                                                                                |
|     |           | 2                                  |                                                                                |
|     | .         | 4                                  |                                                                                |
|     | .         | 6                                  |                                                                                |
*   0 * **.       * 11  12  13  14  16  17  20  8   9   * Conventions         Organization        Sentence Formation  Style              *
|     | ****.     | 15  18  3   5                      |                                                                                |
|     | *****.    | 19  21  7                          |                                                                                |
|     | ********* |                                    |                                                                                |
|     | *******.  |                                    |                                                                                |
|  -1 + *****.    +                                    +                                                                                |
|     | ****      |                                    |                                                                                |
|     | *.        |                                    |                                                                                |
|     | *.        |                                    |                                                                                |
|     | .         |                                    |                                                                                |
|  -2 + .         +                                    +                                                                                |
|     | .         |                                    |                                                                                |
|     | .         |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     | .         |                                    |                                                                                |
|  -3 + .         +                                    +                                                                                |
|     | .         |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     | .         |                                    |                                                                                |
|     |           |                                    |                                                                                |
|  -4 +           +                                    +                                                                                |
|     |           |                                    |                                                                                |
|     | .         |                                    |                                                                                |
|     |           |                                    |                                                                                |
|     |           |                                    |                                                                                |
|  -5 +           +                                    +                                                                                |
|     |   Easy    |                                    |                                                                                |
|     | to rate   |                                    |                                                                                |
|     | accurately|            Inaccurate              |                 Easy to rate accurately                                        |
|-----+-----------+------------------------------------+--------------------------------------------------------------------------------|
|Logit| * = 8     |           Raters                   |                         Domains                                                |
+-------------------------------------------------------------------------------------------------------------------------------------+
```

Figure 14. *Georgia Writing Results: Interactions between Rater Accuracy and Domains based on MFR Model for Rater Accuracy (Model II)*



Note. Values of the *t*-statistic shown here are tests of the hypothesis that there is no interaction between rater accuracy and difficulty to rate domains accurately. Test statistic values higher than *t* = +2.00 suggest that the rater was more accurate than expected for a domain, based on the average measure for the domain across the raters. Test statistic values lower than *t* = -2.00 suggest that the rater was less accurate than expected for a domain.

Figure 15. *Georgia Writing Results: Interactions between Rater Accuracy and Student Gender based on MFR Model for Rater Accuracy (Model II)*



Note. Values of the *t*-statistic shown here are tests of the hypothesis that there is no interaction between rater accuracy and student gender. Test statistic values higher than *t* = +2.00 suggest that the rater was more accurate than expected for a subgroup of students, based on the average measure for the subgroup across the raters. Test statistic values lower than *t* = -2.00 suggest that the rater was less accurate than expected for a subgroup of students.

Figure 16. *Operating Characteristic Functions (OCFs)/Item Response Functions based on Mokken scaling*

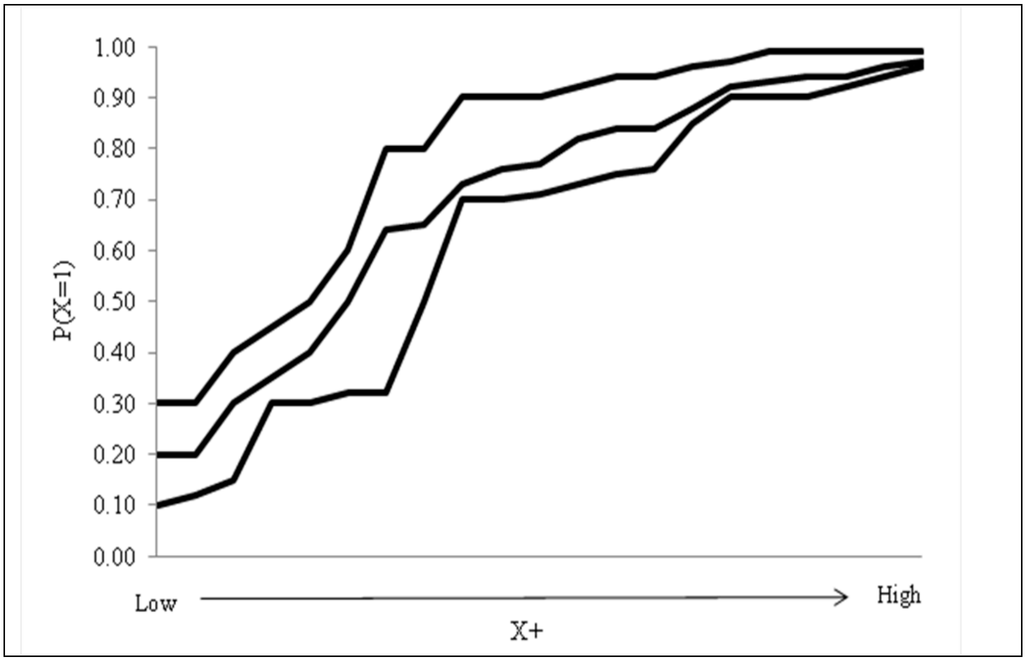
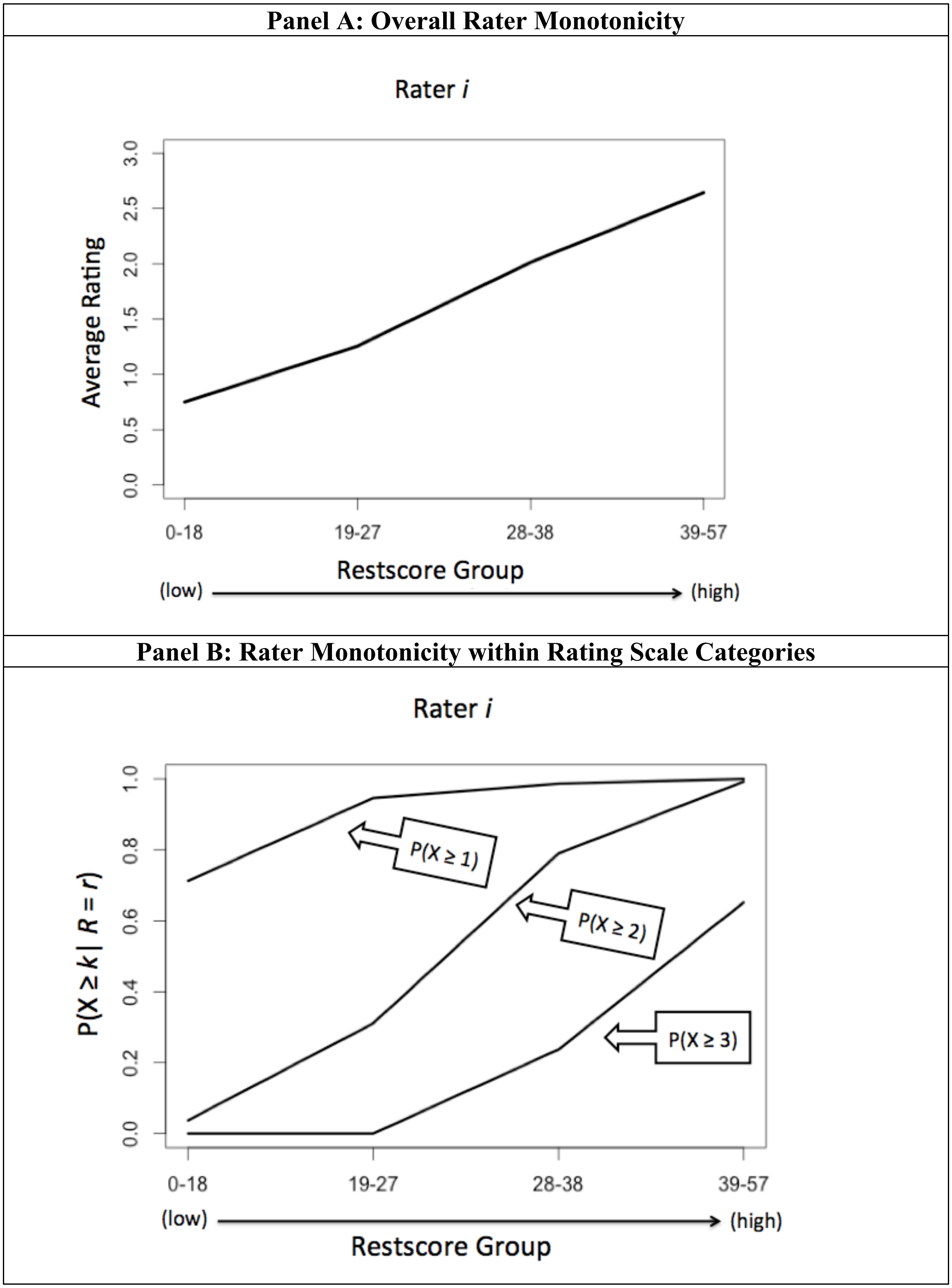


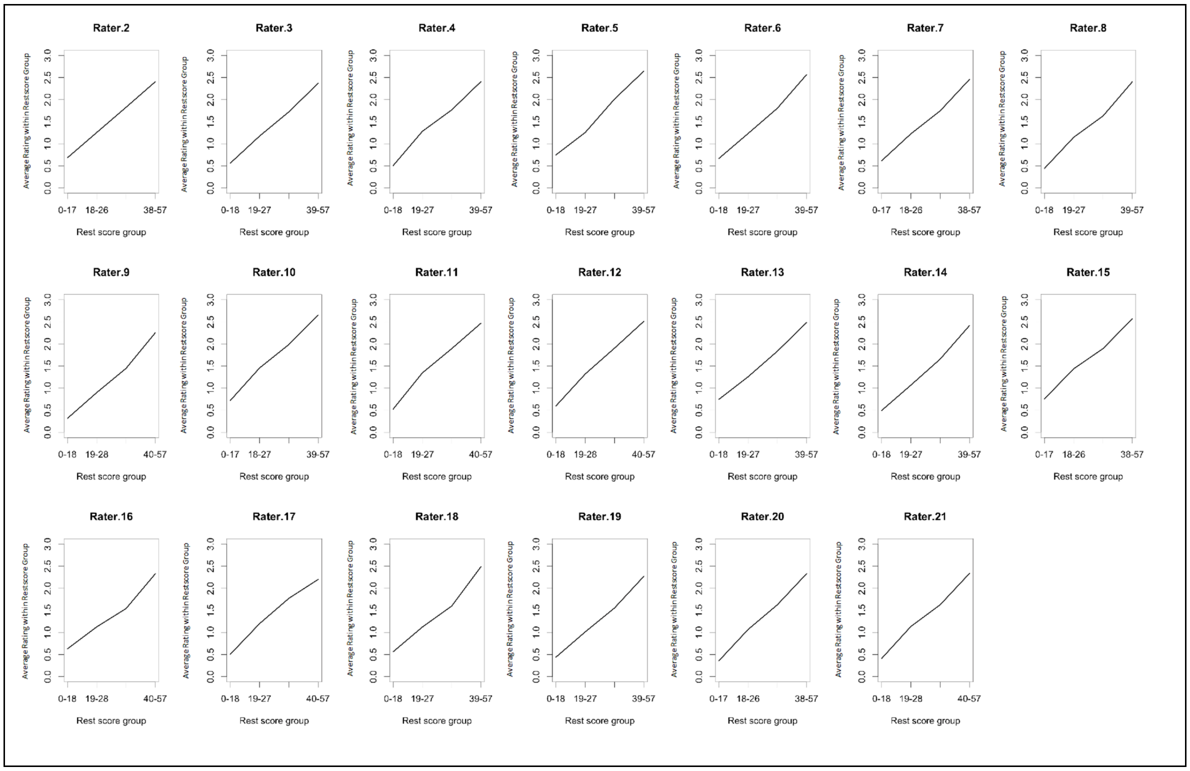
Figure 17. *Illustrative Rater Monotonicity Plots*



Panel A: Overall Rater Monotonicity

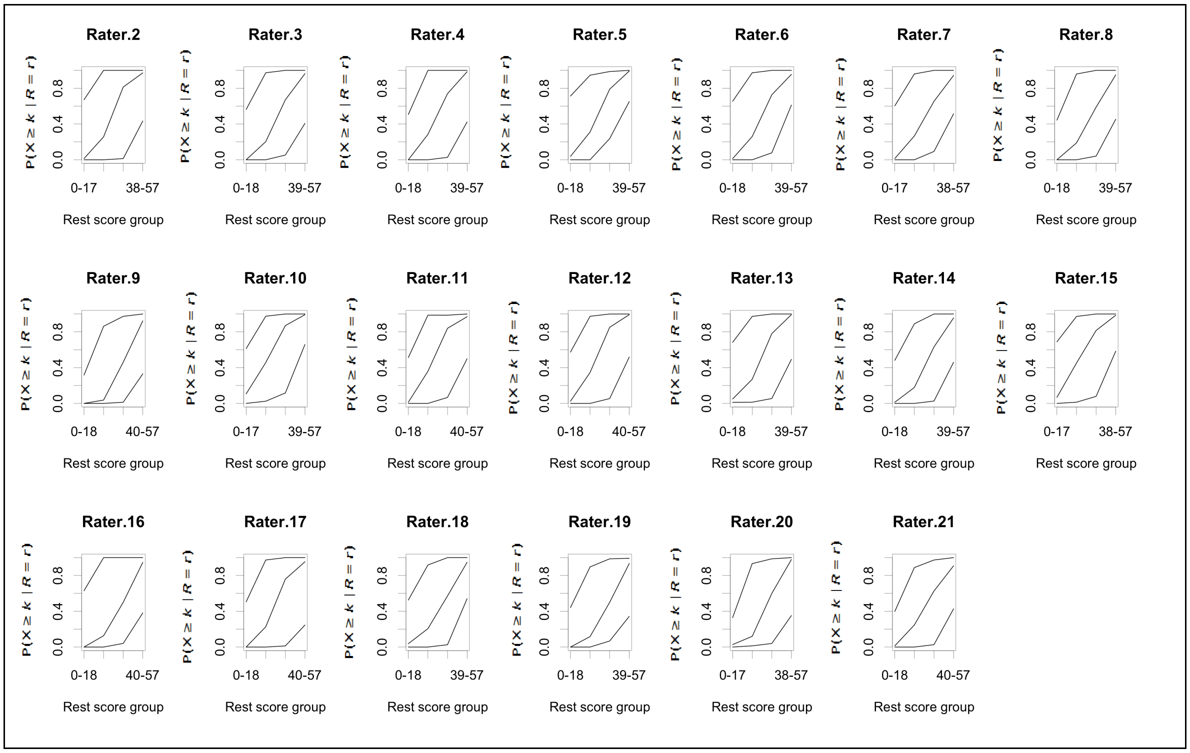
Panel B: Rater Monotonicity within Rating Scale Categories

Figure 18. *Georgia Writing Results: Overall Rater Monotonicity based on MH-R Model (Model III)*



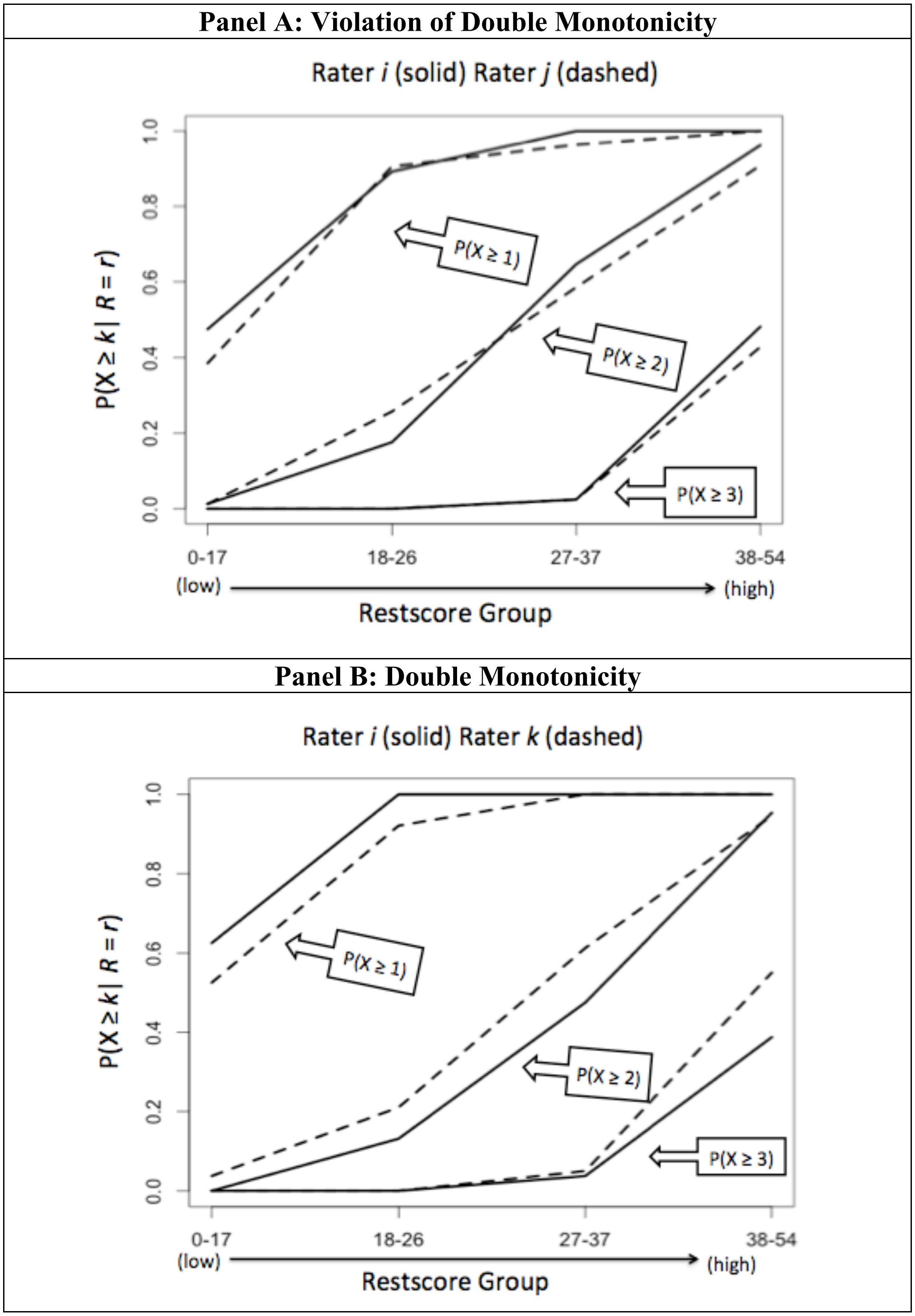
Note. Each plot describes monotonicity for a single operational rater. Restscore groups are plotted along the *x*-axis, and average ratings are plotted along the *y*-axis. Rater monotonicity is implied when average ratings increase as restscores increase.

Figure 19. *Georgia Writing Results: Rater Monotonicity within Rating Scale Categories based on MH-R Model (Model III)*



Note. Each plot describes monotonicity for a single operational rater. Restscore groups are plotted along the *x*-axis, and the cumulative probability for a rating in category *k* is plotted along the *y*-axis. The three lines represent the three meaningful category response functions for the four-category rating scale. The lowest line is the cumulative probability for a rating in Category 3, the middle line is the cumulative probability for a rating in Category 2, and the highest line is the probability for a rating in Category 1.

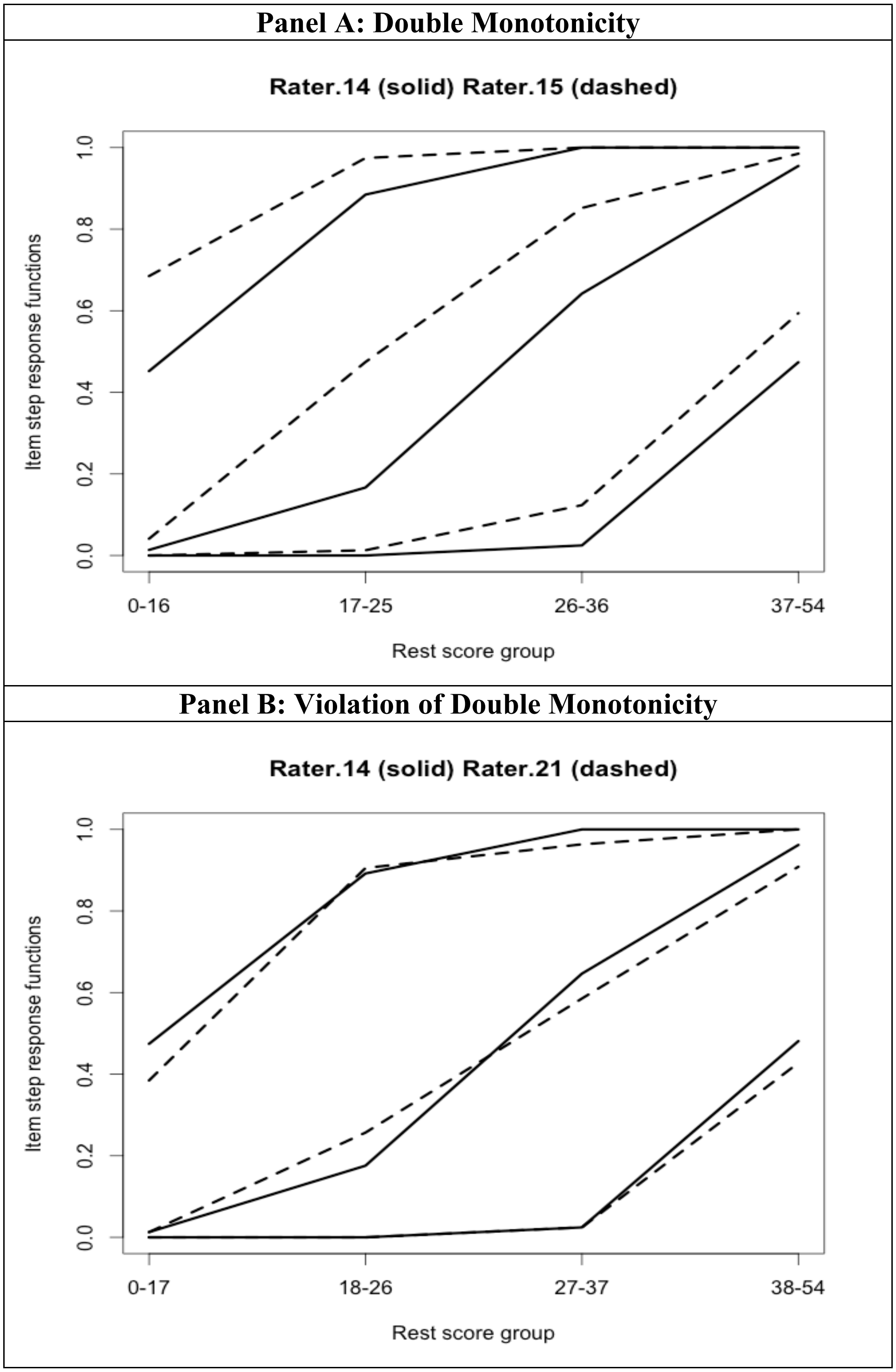
Figure 20. *Illustrative Rater Double Monotonicity Plots via the Restscore Method (Model IV)*



Panel A: Violation of Double Monotonicity

Panel B: Double Monotonicity

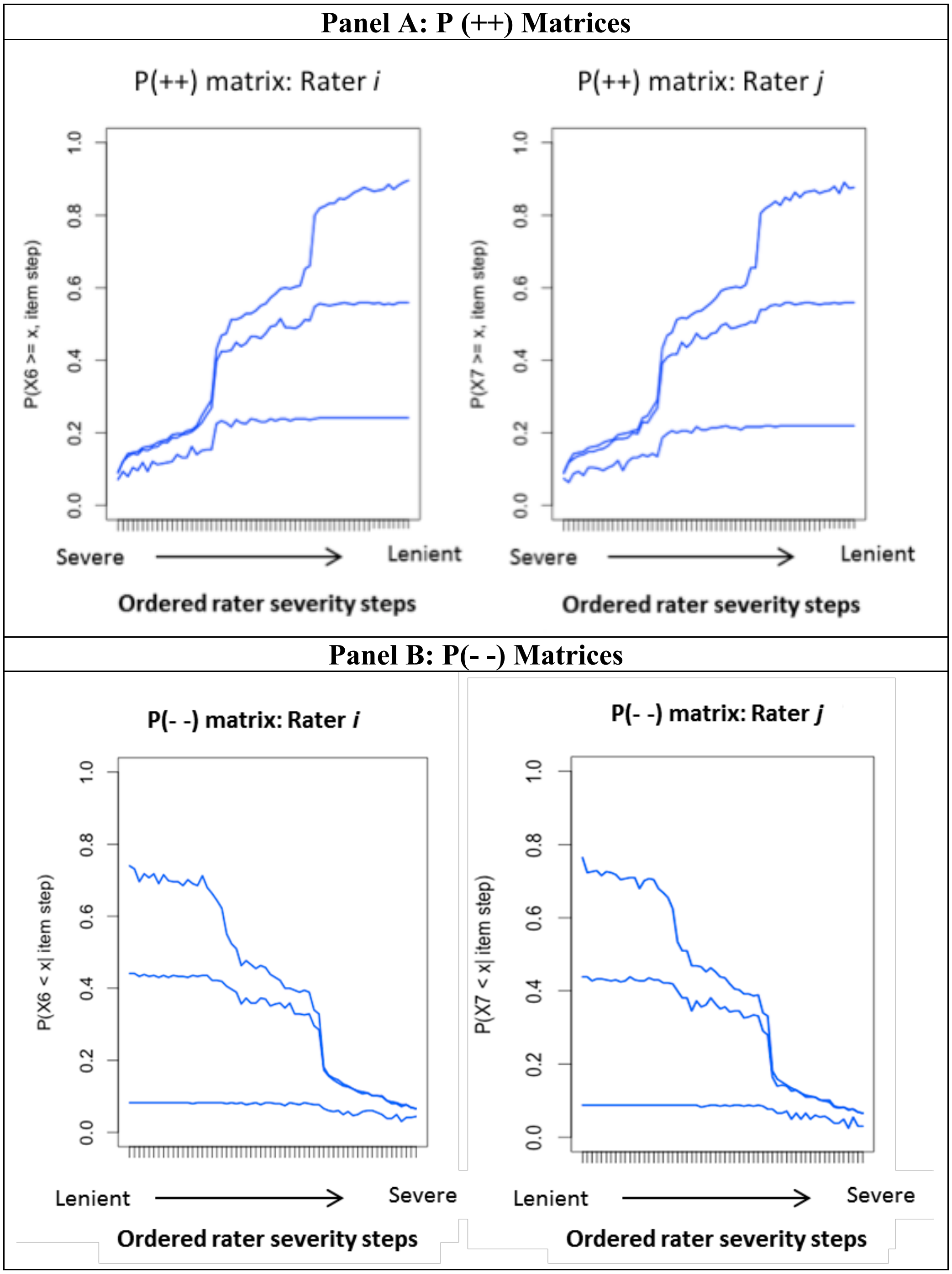
Figure 21. *Georgia Writing Results: Double Monotonicity for Two Rater Pairs Involving Rater 14 via the Restscore Method (Model IV)*



Panel A: Double Monotonicity

Panel B: Violation of Double Monotonicity

Figure 22. *Illustrative Rater Double Monotonicity Plots via the P Matrix Method (Model IV)*



Note. The *x*-axis displays the 57 rating scale category steps (20 raters minus rater of interest = 19 raters, 3 rating scale category steps each), ordered from severe raters to lenient raters, against which the joint probability for a ‘1’ rating on rating scale category step is plotted. The highest line represents the joint probability for passing the step from category 0 to category 1, the middle line represents the probability for passing the step from category 1 to category 2, and the lowest line represents the joint probability for passing the step from category 2 to category 3.

Figure 23. *Georgia Writing Results: Rater Double Monotonicity via the P Matrix Method (Model IV): P(+,+) Results*

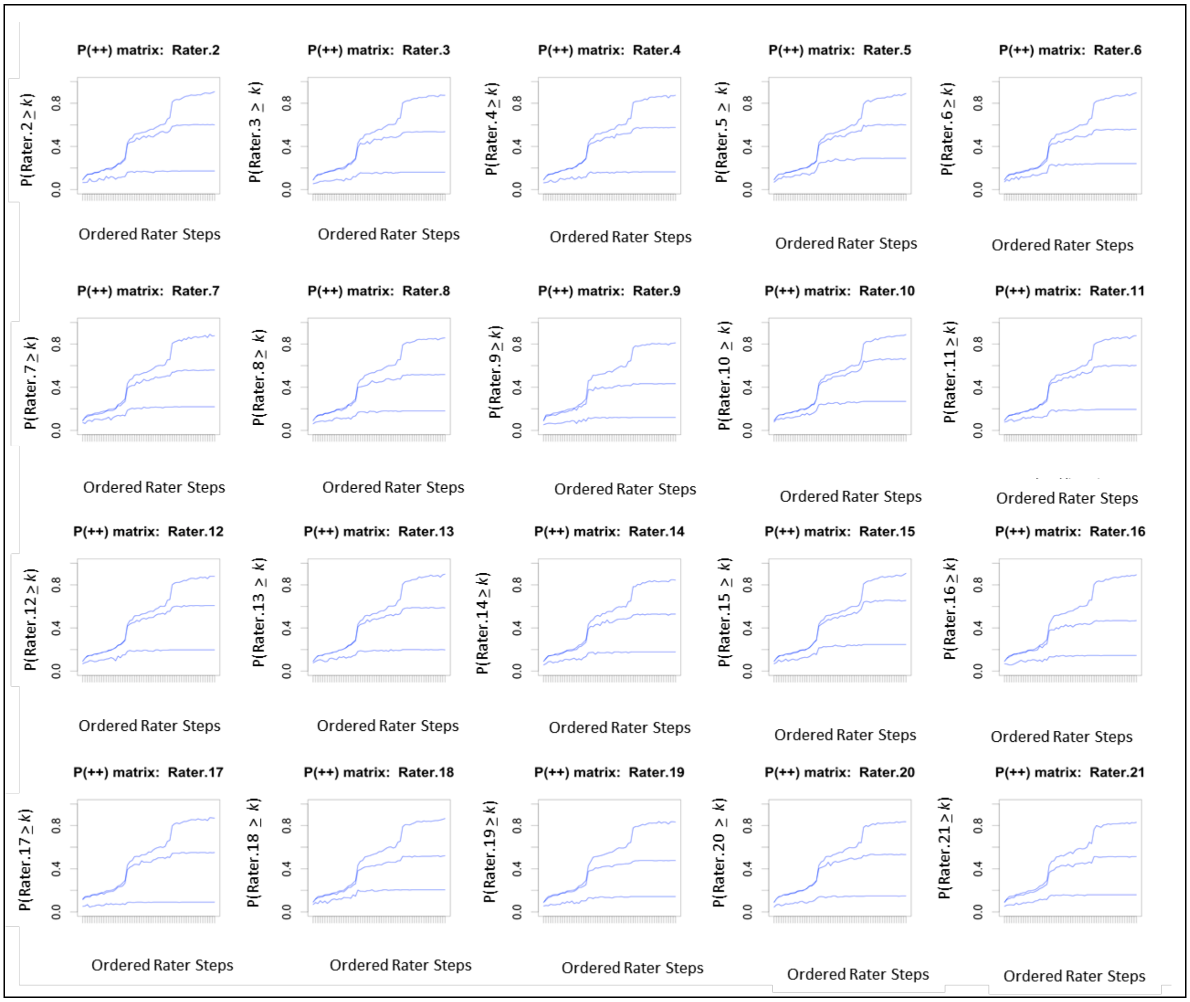


Figure 24. *Georgia Writing Results: Rater Double Monotonicity via the P-Matrix Method (Model IV): P(–, –) Results*

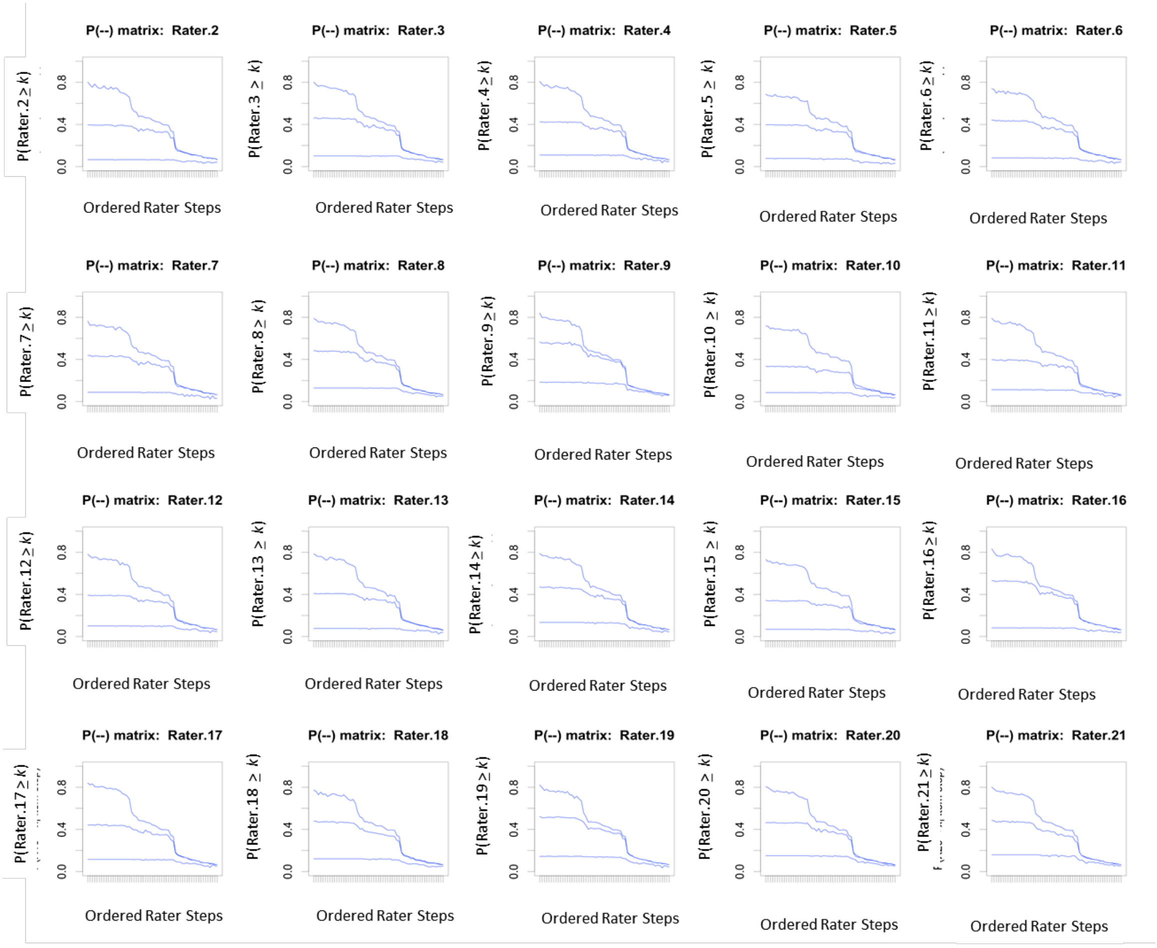
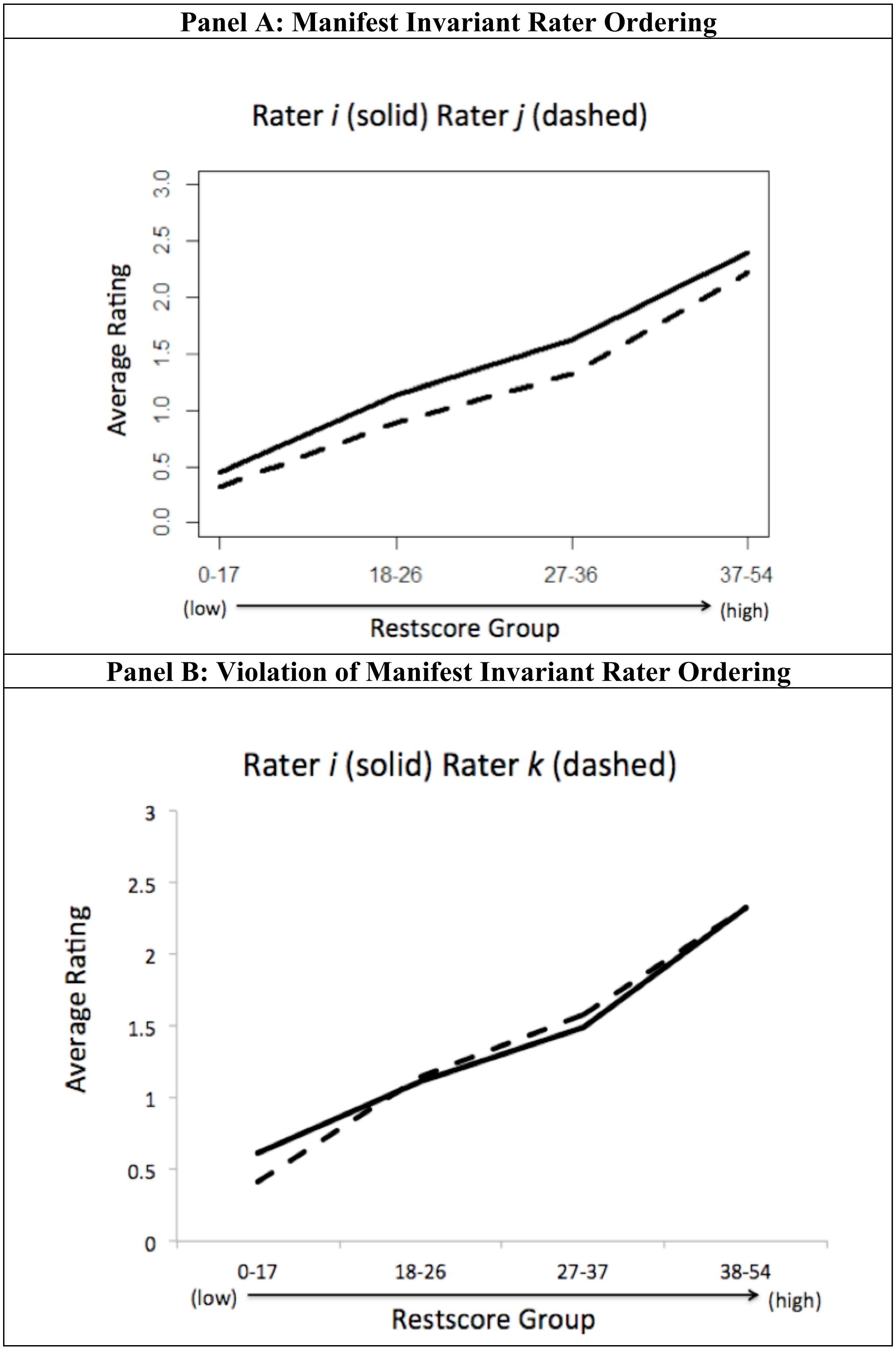


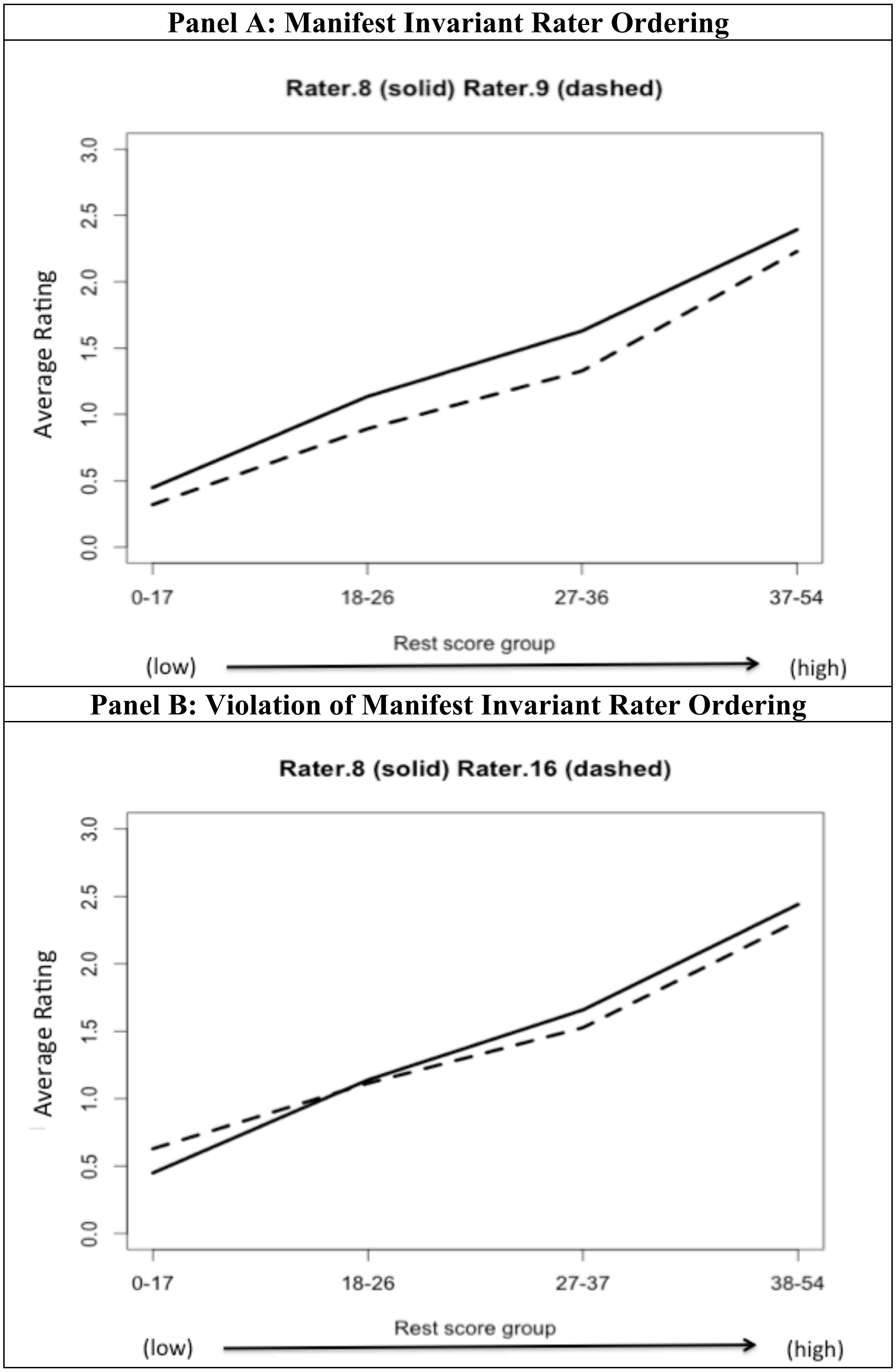
Figure 25. *Diagnostic Plots for Manifest Invariant Rater Ordering*



Panel A: Manifest Invariant Rater Ordering

Panel B: Violation of Manifest Invariant Rater Ordering

Figure 26. *Georgia Writing Results: Manifest Invariant Rater Ordering Plots for Two Rater Pairs Involving Rater 8*



Panel A: Manifest Invariant Rater Ordering

Panel B: Violation of Manifest Invariant Rater Ordering

Figure 27. *Indices of Orientation with the Latent Variable*

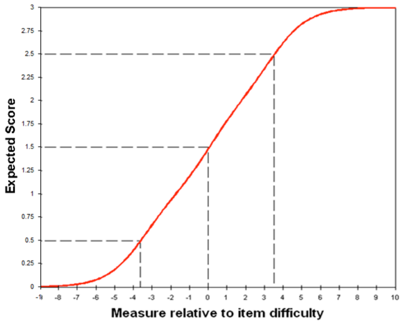
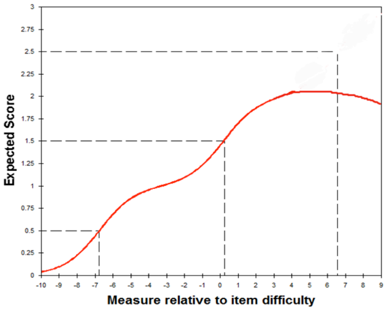
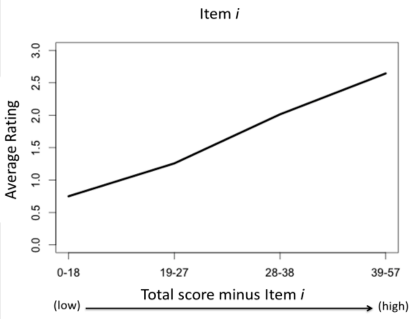
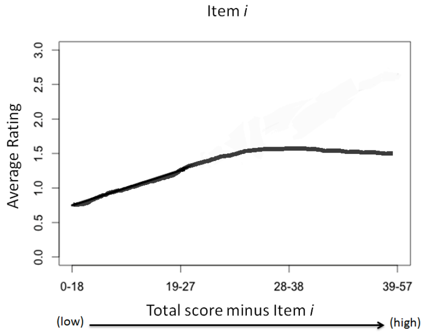
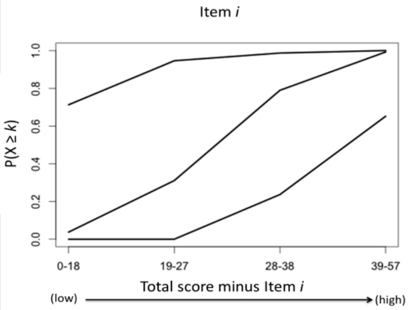
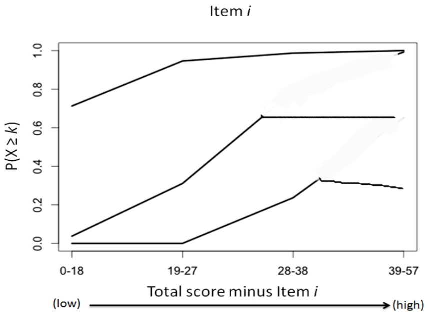
Guideline		No Violation(s)	Violation(s)
Rasch	A. Expected ratings increase monotonically across the latent variable		
Mokken	B. Average ratings increase across rest scores		
	C. Cumulative category probabilities increase across rest scores		

Figure 28. *Indices of Category Precision*

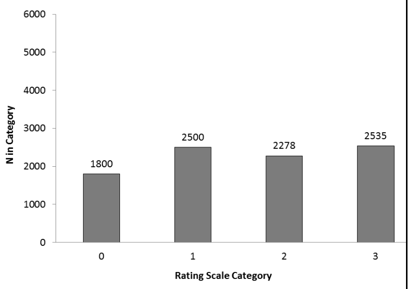
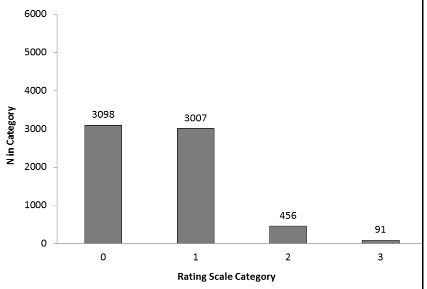
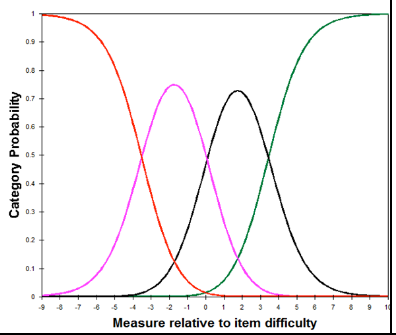
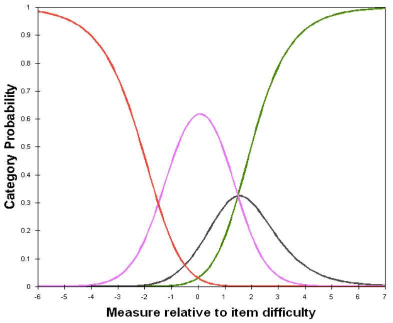
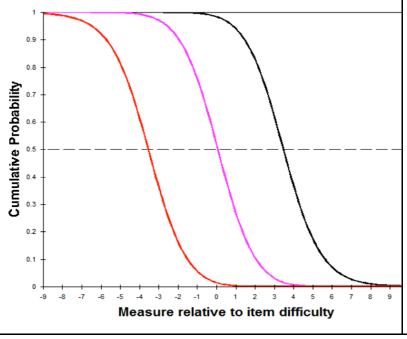
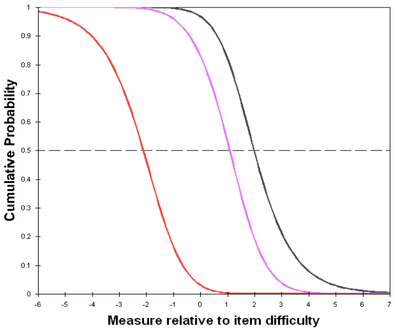
<table>
<tr><th colspan="2">Guideline</th><th>No Violation(s)</th><th>Violation(s)</th></tr>
<tr><td rowspan="4">Rasch</td><td>A. Normal or uniform distribution of ratings across categories</td><td> </td><td> </td></tr>
<tr><td>B. |Difference| between category coefficient locations (logit scale) is somewhat even across categories</td><td><table><tr><th>Category</th><th>δ_{ij}</th><th>|Difference|</th></tr><tr><td>0</td><td></td><td></td></tr><tr><td>1</td><td>-3.51</td><td></td></tr><tr><td>2</td><td>0.07</td><td>3.58</td></tr><tr><td>3</td><td>3.44</td><td>3.37</td></tr></table></td><td><table><tr><th>Category</th><th>δ_{ij}</th><th>|Difference|</th></tr><tr><td>0</td><td></td><td></td></tr><tr><td>1</td><td>- .76</td><td></td></tr><tr><td>2</td><td>1.5</td><td>3.26</td></tr><tr><td>3</td><td>1.49</td><td>0.01</td></tr></table></td></tr>
<tr><td>C. Multimodal category response functions</td><td> </td><td> </td></tr>
<tr><td>D. Conditional probability curves are distinct and evenly spaced along the logit scale</td><td> </td><td> </td></tr>
</table>

Figure 28, continued

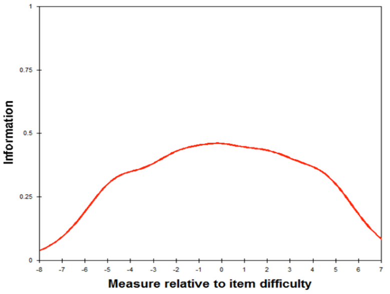
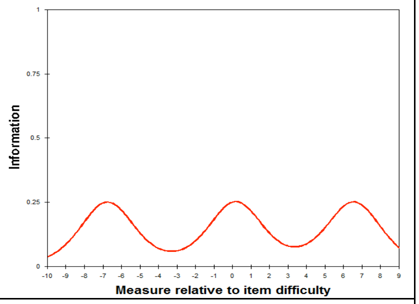
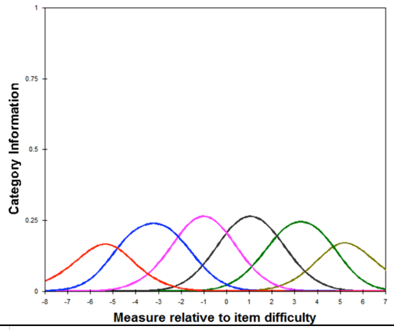
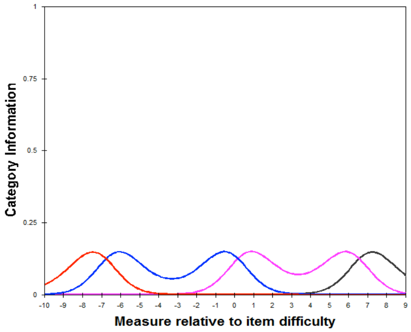
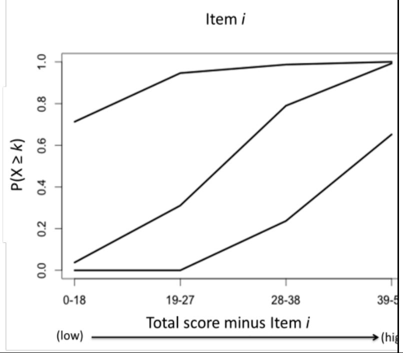
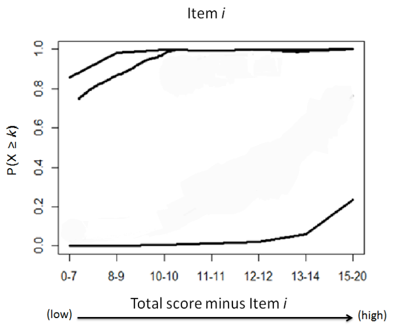
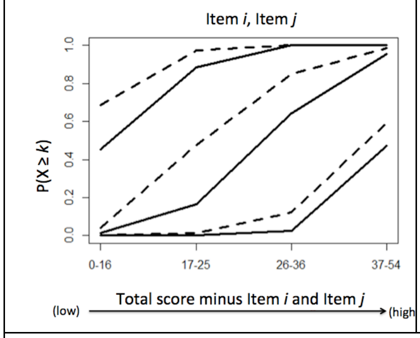
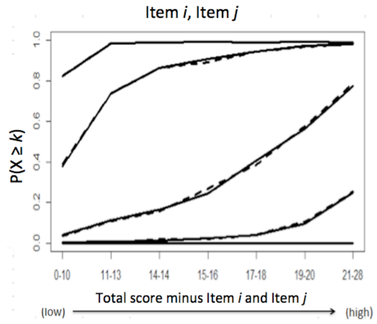
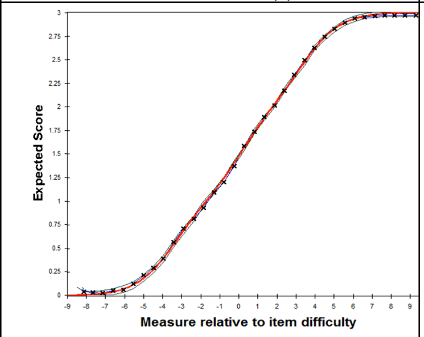
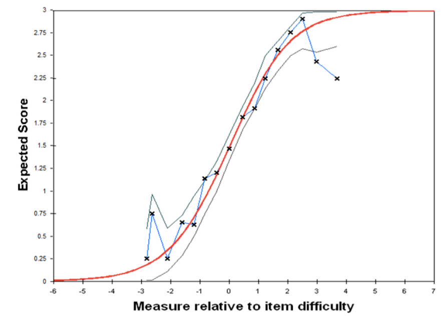
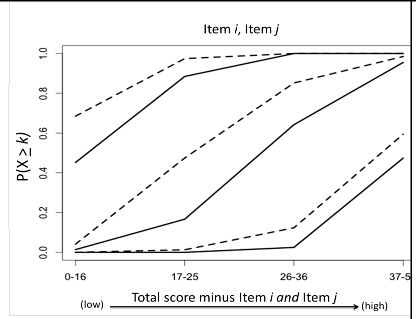
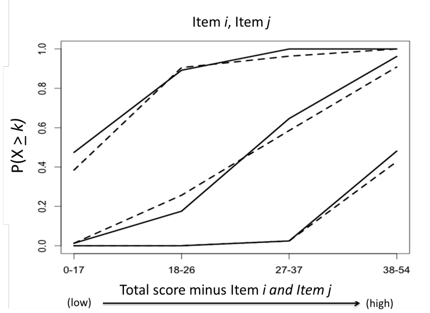
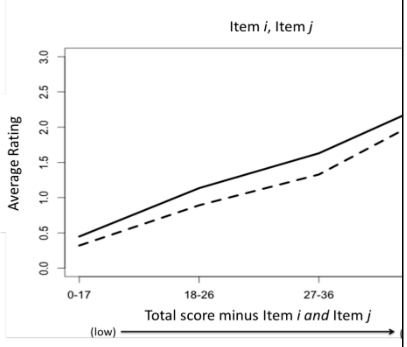
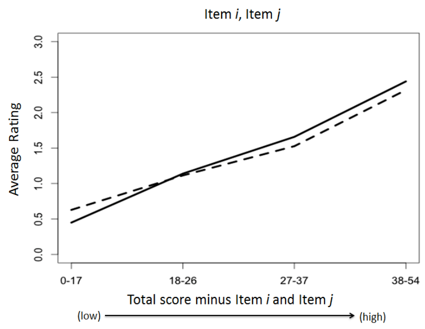
	Guideline	No Violation(s)	Violation(s)
Rasch	E. Smooth item information function		
	F. Smooth category information functions		
Mokken	G. Cumulative category probabilities do not overlap *within* items		
	H. Cumulative category probabilities do not overlap *across* items		
		Item *i* (- - -), Item *j* (—)	

Figure 29. *Indices of Model-data fit*

<table>
<tr><th colspan="2">Guideline</th><th>No Violation(s)</th><th>Violation(s)</th></tr>
<tr><td rowspan="2"><b>Rasch</b></td><td>A. Close match between observed and expected score ogive</td><td> </td><td> </td></tr>
<tr><td>B. Outfit <i>MSE</i> statistics for categories are near their expected value (~1.00)</td><td>
<table><tr><th>Category</th><th>Outfit <i>MSE</i></th></tr><tr><td>0</td><td>0.90</td></tr><tr><td>1</td><td>0.95</td></tr><tr><td>2</td><td>0.90</td></tr><tr><td>3</td><td>0.97</td></tr></table>
</td><td>
<table><tr><th>Category</th><th>Outfit <i>MSE</i></th></tr><tr><td>0</td><td>1.30</td></tr><tr><td>1</td><td>1.20</td></tr><tr><td>2</td><td>1.50</td></tr><tr><td>3</td><td>0.76</td></tr></table>
</td></tr>
<tr><td rowspan="4"><b>Mokken</b></td><td>C. Item scalability coefficients (H_i) suggest scalable items ($\geq \sim 0.3$)</td><td>$H_i = 0.56$</td><td>$H_i = 0.24$</td></tr>
<tr><td>D. Cumulative category probabilities do not intersect across items</td><td> </td><td> </td></tr>
<tr><td></td><td colspan="2">Item <i>i</i> (- - -), Item <i>j</i> (—)</td></tr>
<tr><td>E. Manifest invariant item ordering is observed</td><td> </td><td> </td></tr>
</table>

Appendix A: Rater Scalability

Appendix A is a supplement to the presentation of rater scalability based on Mokken scale analysis in Chapter 4. Part I demonstrates the computation of rater pair scalability for dichotomous ratings using the covariance/maximum covariance method and the error count method. Part II demonstrates the computation of rater scalability coefficients for polytomous ratings.

I. Rater Pair Scalability for Dichotomous Ratings

a. Covariance/maximum covariance method

The scalability coefficient for two raters can be calculated using the covariance formula for binary variables. This method is described by Sijtsma and Molenaar (2002), p. 51 – 52. In this example, ratings assigned by two raters are used to illustrate the calculation of a rater pair scalability coefficient. First, the dichotomized ratings are given in Table A1:

Table A1. *Dichotomized ratings assigned by Rater 1 and Rater 3 to 32 essays:*

Person	Rater 1	Rater 3
1	1	0
2	1	1
3	1	1
4	0	1
5	1	0
6	1	1
7	1	1
8	1	0
9	0	0
10	1	1
11	1	1
12	1	1
13	1	1
14	1	1
15	1	1
16	1	1
17	1	1
18	0	1
19	1	1
20	1	0
21	0	0
22	1	1
23	1	1
24	0	0

Table A1, continued:

Person	Rater 1	Rater 3
25	1	0
26	1	1
27	0	0
28	0	0
29	1	0
30	1	0
31	0	1
32	1	1
Rater Mean	0.75	0.63
SD	0.44	0.49

Step 1: Obtain the 2 x 2 table for the rater pair.

Table A2. *Observed joint frequencies of Rater 1 and Rater 3*

	Rater 3 = 0	Rater 3 = 1	*Total*
Rater 1 = 0	5	3	8
Rater 1 = 1	7	17	24
Total	12	20	32

Step 2. Calculate the observed covariance between the two raters. Use the proportions of ratings in Table A2 to calculate the covariance between Rater 1 and Rater 3.

Table A3. *Rating proportions*

	Rater 3 = 0	Rater 3 = 1	*Total*
Rater 1 = 0	0.15625	0.09375	0.25
Rater 1 = 1	0.21875	0.53125	0.75
Total	0.375	0.625	1.00

$\mathrm{Cov}(X_i, X_j) = P_{ij} - P_i P_j$
Cov(R1, R3) = (R1=1, R3=1) – (R1 = 1 * R3 = 1)
Cov(R1, R3) = 0.53125 – 0.46875 = 0.0625

Step 3. Identify the error cell. Using the observed joint frequencies, identify the error cell. In this example, Rater 1 is more lenient (M= 0.75) than Rater 3 (M = 0.63), so any observations of (R_1,R_3 = 0, 1) are defined as errors. There are 3 observed errors in the top right cell (9% of the observed ratings are errors).

Step 4: Change the observed error frequency to 0. Keeping the row and column marginals fixed, change the frequency of the error cell to 0.

Table A4. *Joint frequencies of Rater 1 and Rater 3 with no errors*

	Rater 3 = 0	**Rater 3 = 1**	***Total***
Rater 1 = 0	8	**0**	8
Rater 1 = 1	4	20	24
Total	12	20	32

Step 5. Obtain the maximum covariance using the new 2 x 2 table. Calculate the covariance between Rater 1 and Rater 3 with no errors.

Table A5. *Rating proportions with no errors*

	Rater 3 = 0	**Rater 3 = 1**	***Total***
Rater 1 = 0	0.25	0.00	0.25
Rater 1 = 1	0.125	0.625	0.75
Total	0.375	0.625	1.00

Cov(R1, R3) = 0.625 – (.75*.625) = **0.15625**

Step 6. Find the rater pair scalability coefficient using the formula H_{ij} = Cov(R_i, R_j)/Cov_{max}(R_i, R_j)

$$H_{R1R3} = 0.0625/0.15625 = \mathbf{0.40}$$

Observed and Expected Error Count Method

The scalability coefficient for two raters can also be calculated using counts of observed and expected errors. This method is described by Sijtsma and Molenaar (2002), p. 53 - 54.

Step 1: Obtain the 2 x 2 table for the rater pair.

Table A6. *Observed joint frequencies of Rater 1 and Rater 3*

	Rater 3 = 0	**Rater 3 = 1**	***Total***
Rater 1 = 0	5	3	8
Rater 1 = 1	7	17	24
Total	12	20	32

Step 2. Identify the error cell. Using the observed joint frequencies, identify the error cell. In this example, Rater 1 is more lenient ($M = 0.75$) than Rater 3 ($M = 0.63$), so any observations of (R_1, R_3 = 0, 1) are defined as errors.

Step 3: Find the observed frequency of errors, *F*. There are 3 observed errors in the top right cell.

Step 4: Find the expected error frequency, *E*, under the null model of independence. Find the expected frequency for the error cell using the following formula

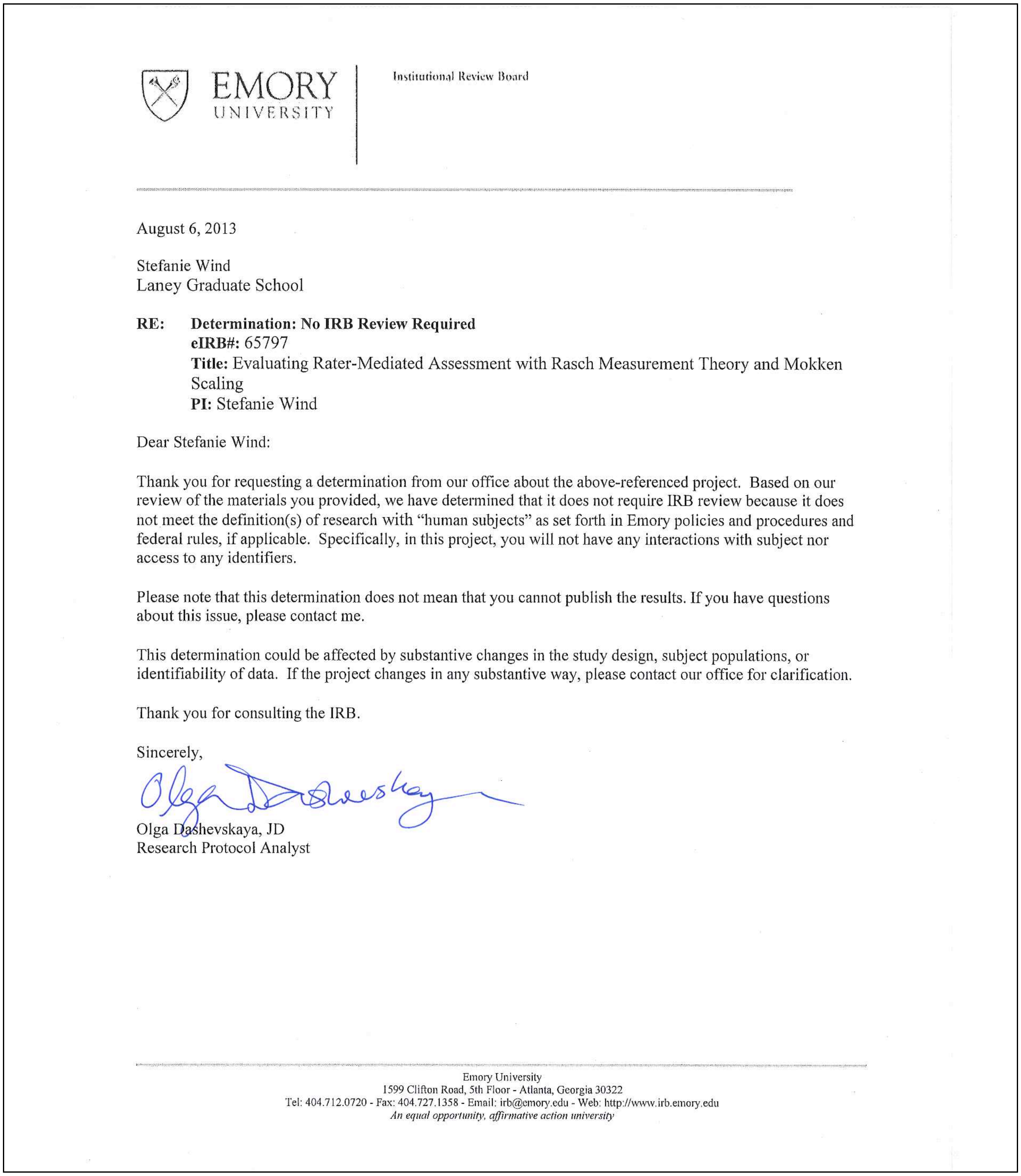
Expected error (Rater *i*, Rater *j*) = (row sum *i*) (column sum *j*) / Total

E (R1, R3) = (8 * 20)/32 = 160/32 = **5**

Step 5: Calculate the rater pair scalability coefficient using the formula $H_{ij} = 1 - (F/E)$

$H_{ij} = 1 - (F/E)$
$H_{R1R3} = 1 - (3/5) = 1 - 0.6 =$ **0.40**

Appendix B: IRB Approval for Alaska Writing Data



EMORY
UNIVERSITY

Institutional Review Board

August 6, 2013

Stefanie Wind
Laney Graduate School

RE: **Determination: No IRB Review Required**
eIRB#: 65797
Title: Evaluating Rater-Mediated Assessment with Rasch Measurement Theory and Mokken Scaling
PI: Stefanie Wind

Dear Stefanie 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 with "human subjects" as set forth in Emory policies and procedures and federal rules, if applicable. Specifically, in this project, you will not have any interactions with subject nor access to any identifiers.

Please note that this determination does not mean that you cannot publish the results. If you have questions about this issue, please contact me.

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,

Olga Dashevskaya

Olga Dashevskaya, JD
Research Protocol Analyst

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