October 2018

Examining the Effects of Changes in Automated Rater Bias and Variability on Test Equating Solutions

Michelle Boyer

EXAMINING THE EFFECTS OF CHANGES IN AUTOMATED RATER BIAS AND VARIABILITY ON TEST EQUATING SOLUTIONS

A Dissertation Presented

by

MICHELLE BOYER

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2018

College of Education

EXAMINING THE EFFECTS OF CHANGES IN AUTOMATED RATER BIAS AND VARIABILITY ON TEST EQUATING SOLUTIONS

A Dissertation Presented

By

MICHELLE L. BOYER

Approved as to style and content by:

_____
Lisa Keller, Chairperson

_____
Craig Wells, Member

_____
Elizabeth Harvey, Member

_____
Richard Patz, Member

_____
Jennifer Randall
Associate Dean of Academic Affairs
College of Education

DEDICATION

To Dan, Sean, and Meagan: You are my world.

ACKNOWLEDGMENTS

I want to thank Lisa Keller for sharing your expertise, wisdom, and thoughtfulness, not only through your guidance on this project, but throughout my time and coursework at Umass. I am so very lucky to have had this chance to work closely with you, to learn from your guidance and from your example. I would also like to thank my Umass Committee members, Craig Wells and Lisa Harvey. Your thoughtful review and comments have helped shape this work into a more rigorous and useful study than it was originally conceived. I am so grateful for your support.

I want to especially thank Rich Patz, not only for sharing your expertise and mentorship as a member of my dissertation committee, but for pushing me at every point to deepen my learning and to seek to increase the value of this contribution to the field. I also do not want to miss this opportunity to thank you for the many years that you provided me with an extraordinary number of opportunities, not just for learning, but for outright adventures in psychometrics. These experiences, and your example strongly influenced my decision to pursue a career in psychometrics, and working with you on my dissertation has helped to prepare me well to continue on this path.

To Ron Hambleton, Scott Monroe, Jennifer Randall, and Steve Sireci: The last four years have been incredibly personally and professionally rewarding for me. I have each of you to thank for that. I have enjoyed every one of your courses and our collaborations, and I aspire to take the knowledge, skills, professionalism, and ethics that you have shared, forward into the profession in ways that I hope will make you proud. For certain, these same qualities that you have demonstrated have helped to guide and shape this project.

A special thank you goes to Craig Mills for your personal recommendation to the Umass faculty, and for allowing me to make my studies a part of my job. Without your kind and generous actions, this work would not have been possible. You believed in me where I did not.

My first years working in testing were unexpectedly exciting. Having imagined myself as a professor of political science somewhere on the East Coast, I could not have found myself more shocked to be in Anne Fitzpatrick's office, in Monterey, CA, as she helped me to study Hambleton, Swaminathan, and Rogers (1991). I recall telling her that I could not imagine running all those calibrations without understanding the underlying concepts and methods. That was eighteen years ago, and possibly I do understand now what I was doing then. Thank you, Anne, for generously sharing your time and mentorship to help me get started on this path.

Finally, I want to thank Ross Green. You didn't want to hire me, but you felt compelled to do so after receiving my application twice. As I sat in that memorable green chair in your office during my first interview you told me that, if I cared about education, this was a place I could make a difference. To this day, when I am confronted with a challenge, I ask myself, "How would Ross have handled this?" And I still hope that I can make that difference, Ross.

ABSTRACT

EXAMINING THE EFFECTS OF CHANGES IN AUTOMATED RATER BIAS AND VARIABILITY ON TEST EQUATING SOLUTIONS

SEPTEMBER 2018

MICHELLE BOYER, B.S., ILLINOIS STATE UNIVERSITY

M.A., BOWIE STATE UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Lisa Keller

Many studies have examined the quality of automated raters, but none have focused on the potential effects of systematic rater error on the psychometric properties of test scores. This simulation study examines the comparability of test scores under multiple rater bias and variability conditions, and addresses questions of their effects on test equating solutions. Effects are characterized by a comparison of equated and observed raw scores and estimates of examinee ability across the bias and variability scenarios. Findings suggest that the presence of, and changes in, rater bias and variability affect the equivalence of total raw scores, particularly at higher and lower ends of the score scale. The effects are shown to be larger where variability levels are higher, and, generally, where more constructed response items are used in the equating. Preliminary findings also suggest that consistently higher rater variability may have a slightly larger negative impact on the comparability of scores than does reducing rater bias and variability under the conditions examined here. Finally, a non-equivalent groups anchor test (NEAT) equating design may be slightly more robust to changes in rater bias and variability than a single group equating design for the bias scenarios investigated.

CONTENTS

Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

CHAPTER I
INTRODUCTION

1.1 Introduction

Demands for assessments that measure the increasingly complex cognitive processes required for high school and college graduates to be successful (Conley, 2014), along with the high cost and slow speed of traditional human scoring processes, has increased pressures on test developers to consider the use of automated scoring. Although experts in the fields applied and computational linguistics, statistics, computer science, natural language processing, cognitive science, and psychometrics have made a great deal of progress in developing automated scoring methods that are capable of matching (and perhaps exceeding) human rater accuracy in essay scoring (Shermis & Hamner, 2012; 2013; Shermis, 2015; Kieftenbeld & Boyer 2017), challenge remain in the development of automated approaches to scoring examinee constructed responses with the same accuracy as traditional human scoring.

As automated methods for constructed response scoring methods are introduced it will be important to understand their impact on the psychometric properties of tests and the statistical comparability of scores across administrations. This understanding may be particularly important under conditions where different automated raters are used to score items that are common across test forms, but is a concern for equating scenarios generally. Where item scores contain changing levels of rater error (due to any source), differences in the inferences that can be made about what examinees know and can do could be present across test administrations. Where the comparability of scores across administrations is targeted, such differences would represent a threat to the validity of the test results. Section 1.2 and 1.3 provide a brief overview of automated scoring, some definitions, and the psychometric context that motivates this study.

1.2 Automated Scoring: Important Definitions and Background

Broadly, automated scoring is any automated process for scoring examinee responses to items on tests. Examinee responses are input into the process, and scores and other types of feedback are output. Expectations for examinee responses for each item or collection of items are typically defined by a set of rules that may be as simple as providing an answer key for the correct lettered or numbered response among several other possible responses; or the rules may be defined in a scoring rubric which details elements of responses expectations to varying degrees of specificity. Setting aside a consideration of the various human interventions that might exist in such processes for now, there are many methods that might be used to produce test scores for examinees depending on both item type, and the structure of the expectations for examinee responses.

Although not the focus of this study, multiple-choice scoring based on answer keys is perhaps the most basic and common form of automated scoring and represents its earliest form. Examinees select one or more responses from a set of options on paper response documents that are scannable, or in a computer-based environment. Whether through optical mark recognition of scannable documents or through computer capture of key strokes, examinee responses are matched to answer keys by a straight forward process of automatically matching the answer key to examinee responses.

Similarly, more recent item types have been developed that rely on various pre-specified rules to score examinee responses. What each of these item types share is that they were made possible through ongoing technological advances to make increasingly complex examinee interactions with test content widely available on computers and other devices, such as tablets,

cell phones, and notebooks. The technology enhancement may exist in the item stem as a means to access content, or at the response level to allow response manipulations that demonstrate applied knowledge, skills, and abilities; or the enhancement may exist in both the stem and response. Examples of items that have enhanced response possibilities include those that ask examinees to identify correct responses by "dragging and dropping" one or more response options to the correct on-screen position, items that ask examinees to match information in a table, and items that require other examinee manipulations to "construct" a response within some level of constraint. The point here is that, although much work remains to develop appropriate and meaningful rules to apply to examinee responses, responses to these item types are, by design, constrained in ways that allow the automated process to apply a precise set of rules to score each examinees response. This type of automated scoring is most often applied in computer-based environments and is commonly referred to as rule-based scoring.

Automated scoring of fully constructed responses, however, has the added challenge of appropriately applying scoring rules to examinee responses that are not constrained in controllable and fully predictable ways. The theoretically infinite number of ways an examinee might approach a response to such an item are not addressable through the creation of rules for every possible scenario, and have typically fallen in the domain of traditional human scoring. The approach to automated scoring for this type of item is more difficult and requires a process that can handle the variety of responses that might be provided by examinees. The approaches that have been used are fundamentally different than applying precise rules or keys for item with constrained response expectations. The notion of a scoring rubric still applies, but it is in the application of the rubric to the examinee responses where the methods diverge from typical multiple-choice and rule-based processes.

Automated scoring methods for constructed response items are commonly referred to as artificial intelligence (AI) methods. For the purposes of this study, however, the more general term of "automated scoring" is preferred due to its assessment-specific use of such approaches, and to allow for the future inclusion of alternative methods that do not fall within the realm of traditional AI methods.

The methods that have been developed to address automated scoring of constructed response items often employ a statistical prediction model that relies on scores from expertly trained human raters. Many also use aspects of natural language processing (NLP), such as the methods described in Manning and Schutze (2009) to extract variables ("features") from examinee responses for use in the development and application of the prediction model. The features themselves range from simple traits that can be directly matched or measured such as specified words, di-grams, tri-grams, and sentence length, to deeper features that function as proxies for measuring underlying constructs such as organization and language mechanics for essays, and content expectations for specific domains such as reading, mathematics, and science. Chapter 2 contains a more detailed description of the various methods that have been used and that continue to be developed.

For the purposes of characterizing change in the quality of automated raters, a distinction is made between automated rater methods (or approaches) and models. An automated scoring method is treated here as a general process or approach to scoring, whereas the model is often, although not necessarily, an item-specific statistical model. A method or "approach" to automated scoring can theoretically include as many models as there are there are items to score. Models are empirically developed and applied directly during item level scoring. As it is unlikely

that one statistical model can accurately score all constructed response items, many statistical models may be developed under a single approach or method.

Changes in levels of automated rater error may arise from the appropriateness (or lack thereof) of the method, changes in the method, or changes in the statistical model, or even all three. This can make it difficult to define an AR—is it the method or the model or something else? In practice automated raters are most often referred to by their product names which embody select methods and models, such as e-Rater®, c-Rater™, c-Rater-ML, Intellimetric®, Knowledge Assessment Technologies™ (KAT), and Intelligent Essay Assessor (IEA), but changes in the level of rater error in scoring can occur due to changes within or across such automated rater. For the purpose of this study, then, an automated rater is defined simply by its product name, and it is assumed that changes in rater error may arise from improvements defined at the product, method, or model levels.

1.3 Motivation and Research Question

There may be both intended and unintended consequences to advancing the state of the art in automated scoring. The intended consequences are clear: Develop automated raters that can produce valid and reliable scores, and decrease scoring time and cost. However, an improved scoring model is by definition expected to produce better scores than its predecessor, so as automated scoring methods improve, a potential unintended consequence, may be that such improvements could diminish the comparability of scores across test forms by virtue of applying different or improved methods or statistical models, or even different automated raters. Currently, there are no studies that have been identified that assess the impact of changing levels of rater error on test equating solutions and their outcomes.

There is, therefore, a need to understand the differences in scores produced by different automated raters, and their impact on test results. Score differences due to changing quality of automated raters could be particularly problematic where automated scoring is employed in high stakes assessment programs that demand comparability of scores across administrations, and where testing programs must control rater bias and variation to ensure score comparability, regardless of whether those scores are produced by human or automated raters.

The motivating problem for this study, then, is a consequence of expectations that the state of the art in automated scoring for constructed response items will continue to improve. One direct effect of such improvements can be framed as resultant changes in the level of rater bias and variability, presumably downward. Reductions in bias and variability are highly desirable, however, this would represent a change with the potential to impact test equating solutions in programs that require consistency of scoring practices over time. Assessing the impact of changes in rater bias then, will help to understand when the comparability of test scores might be threatened.

This study will examine conditions of rater bias change and the impact of this change on testing equating solutions to gain a better understanding of whether or not there are circumstances under which special considerations for the use of automated scoring in high stakes operational settings might be warranted. Such considerations may include whether or not to hold the state of the art in automated scoring constant within a testing programs, to consider the use of statistical adjustments during equating, or to perhaps consider the possibility of accepting a lesser equated solution for the sake of using scores with progressively lower rater bias.

1.4 Research Questions

The primary research question that this study addresses is, what are the effects of rater bias and variability on test equating solutions? More specifically, and for the purpose of directly addressing this main question, the following questions are posed: 1) What are the effects of changes in automated rater bias and variability on test equating solutions? 2) Do changes in rater bias and variability change the inferences that can be made about test scores that are intended to be comparable? 3) What is the impact to examinee scores and performance level classifications under these conditions? 4) And, are either common item or common examinee designs more robust to changes in rater bias and variability?

CHAPTER II

BACKGROUND AND REVIEW OF THE LITERATURE

2.1 Overview

Chapter 2 provides the background and research relevant to this study's focus on the impact of automated rater bias on test equating solutions, starting with an overview of test equating designs, properties, and procedures (2.2). The test equating discussion is followed by a description of typical automated scoring methods (2.3), how the quality of automated scoring results are currently framed and measured (2.4), and some possible psychometric implications for the use of automated scoring in operational testing programs (2.5).

2.2 Test Equating Designs, Properties, and Procedures

2.2.1 Equating Designs

Test equating is common where it is desirable to make the same inferences about scores for examinees who take different test forms. The process of test equating is defined by Kolen and Brennan (2014, pg. 2) as, "…a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably." There are two fundamental challenges for treating scores from different forms in this way, 1) test difficulty may vary across forms, and 2) the ability distributions of examinees may differ across groups (Lord, 1980; Kolen & Brennan, 2014). Test equating procedures, in their most general form, involve the development of equating functions based on these differences.

In both classical and modern test theory, a precondition for the use of test equating procedures is that different test forms target the same construct, and are ideally composed of items with similar statistical qualities (Lord, 1980; Kolen and Brennan, 2014; Dorans and

Pommerich, 2009). Fundamentally, equating requires that some common structure across test forms be present in the design to account for differences in test difficulty under multi-form testing conditions. In practice this means that either common items or examinees are used to form the basis for establishing score equivalency.

As defined by Kolen and Brennan (2014), test equating procedures are treated as methods of producing "interchangeable" scores, however, the term interchangeable can be a subject of some debate due to a variety of sources of error that might occur. The measurement model, examinee sampling, and the item level scores that are used to produce the test scale may all contribute error to the equating results. Should the level of error in any one or more of these elements exceed reasonable levels, strict interchangeability of scores may be threatened. As the purpose of this study is to examine the effects of changes in systematic rater error (bias), the term "comparability" of scores is preferred as a less absolute characterization of the relationship between scores on equated forms.

We might want to say, for example that a grade 6 mathematics score of 500 has the same meaning, regardless of which test form an examinee takes. In such a case, form "Y" may be equated to form "X." In some cases, it is additionally desirable to produce scores that are comparable across grades as well as across forms within a grade. In this scenario, tests are equated both within and across grades on a single ("vertical") scale. Continuing the example of a score of 500 on a sixth grade mathematics test form Y, a seventh grade score of 600 (on say, test form Z) for the same individual would be treated as an increase of 100 points on the same scale as the sixth grade score on form X or Y.

Three equating designs are common in the literature, and applied as appropriate to the needs of individual testing programs. Briefly, these designs are, 1) single group design, 2) non-

equivalent-groups anchor test (NEAT), and 3) randomly equivalent groups design. Single group is a common examinees design where, as the name suggests, the same group of examinees takes two or more forms of a test. Practically speaking, this design can be difficult to implement as it represents a significant testing burden for individual examinees taking multiple test forms. Effects from fatigue and order of administration can impact the equating solution. This design is more common in the development of tests where item exposure across test forms is not permitted. Scenarios involving the development of pre- and post-tests to examine treatment effects might require such a design.

A more efficient, and less restrictive design is the NEAT design, where an "anchor" is established by embedding a common set of items in alternate or sequentially developed test forms. Broadly speaking, an equating function is then built based on the differences in scores for the set of common items. Also important in most NEAT designs is that this anchor set of items represent the domain of the construct being measured, and that it have similar statistical properties to the full test (Kolen & Brennan, 2014).

A less common design in operational settings is a randomly equivalent groups design. Here, neither examinees nor items are common across test forms, but examinee groups taking each form are treated as if they were "randomly equivalent." The sampling of examinees under this design can occur either before or after test administration. One way to sample the groups prior to test administration is to spiral multiple test forms throughout the examinee population. Spiraling involves a process, whether the tests are paper or computer based, of distributing or assigning test forms sequentially at the student level. For example, test forms for a classroom of students might be sequenced 1-5. The test administrator hands out the tests 1, 2, 3, 4, 5, to the first 5 students, again for the next 5 students and so on. Although a more experimental sampling

design would be the clear preference over spiraling, such designs are often impractical to implement due to expense and administration complications, such as an extra testing burden on examinees.

Spiraling is also useful under NEAT designs where multiple forms with common items are spiraled to secure roughly equivalent numbers of examinee responses for all items and forms, as well as to establish a stronger link between forms. As spiraling may not result in strictly equivalent groups, it is not often used in operational settings without the presence of common items to strengthen the link. An associated issue is that an accumulation of random error in the equating function is possible where spiraling produces small form level sample sizes that are not strictly equivalent (Braun & Holland, 1982; Kamata and Tate, 2005; Haberman & Dorans, 2009); and further, the accumulation of error across equating solutions varies by the equating method chosen (Keller & Keller, 2011; Keller & Hambleton, 2013).

Another, implementation of a randomly equivalent groups design involves a process of matching samples under observational (after test administration) conditions, based on variables considered to be covariates of examinee ability. These covariates are used to reduce selection bias by statistically matching examinees across the two test forms to control for any systematic differences in the samples used for equating, e.g. through propensity score matching (Rosenbaum and Rubin, 1983). Examples of covariates used to produce matched groups for test equating might include examinee background variables that are known strong predictors of performance such as parental education and other item, testlet, or test scores. In practice, this type of information is often not available and commonly collected demographic vairables are used instead such as gender, ethnicity, grade level…etc. In the case of PSM, such variables are used to create a multivariate composite, or "propensity score" on which examinees are matched

and treated essentially as equivalent groups. Even so, this use of observable covariates might not eliminate the effect of nonrandom group assignments and can lead to bias in subsequent equating procedures (Dorans, 1990; Eignor, Stocking, & Cook, 1990).

This design is used primarily in settings where two conditions exist: equivalent groups sampling under experimental design conditions for two test forms is prohibitive, and no items can be considered to be common on the forms. This situation often arises where non-equivalent groups of examinees take a test in different testing modes, e.g. paper versus computer (Yu, Livingston, Larkin, & Bonett, 2004; Way, Davis, & Fitzpatrick, 2006; Way, Um, Lin, & McClarty, 2007; Way, Lin, & Kong, 2008). Once the randomly equivalent groups are established, they are essentially treated as if the equating design were a single group design.

2.2.2 Equating Properties and Methods

Each of these equating designs might employ any number of statistical procedures to compute an equating function. However, not every equating procedure is appropriate or useful in every equating design. There are three important properties of equating that guide how equating adjustments are made and which procedures are most appropriate under each design. These concepts are, 1) use of the same test specifications, 2) equity, and 3) symmetry (Lord, 1980; Kolen & Brennan, 2014). The same specifications property was described earlier, namely that it is the requirement for all equated tests to measure the same construct domain, and to do so with similar statistical properties across forms such as the range of item difficulty and test reliability. Most often test construction follows precisely defined specifications and blueprints that facilitate the production of multiple test forms to measure performance on the same construct.

Equity is inherent in the goal of equating to provide comparable scores on different test forms. Lord (1980, p. 195) defines *Lord's equity property* of equating as a property that exists when it is a matter of indifference which form an examinee takes. Lord defined this property,

$$G^*[eqy(x)|\tau] = G(y|\tau), \text{ for all values of } \tau \qquad (2.2.1)$$

In equation 2.2.1, τ represents the examinee's true score, x represents a particular observed score on form X, y represents a particular observed score on form Y, eqy is the equating function that converts scores on Y to X, G is the cumulative distribution of scores on Y, and G^* is the cumulative distribution of eqy for the same examinees. Generally, if the equity property holds, the equating procedure should result in the same observed means, standard deviations, and distribution shapes between observed scores Y and converted scores X for any given τ.

Due to the practical difficulty inherent in establishing full equity, Morris (1982) suggested a less strict form of equity, or "first-order equity" which focuses only on the mean of the distribution of scores across forms.

$$E^*[eqy(x)|\tau] = E(y|\tau), \text{ for all values of } \tau \qquad (2.2.2)$$

where E is the expectation operator, and implies that an examinee has the same equated score on forms X and Y. Basically, first order equity holds "to the extent that conditional expected scale scores are similar for the alternate forms" (Kolen & Brennan, 2004, p. 301). Second-order equity holds "to the extent that the conditional standard errors of measurement, after equating, are similar for the alternate forms" (Kolen & Brennan, 2004, p. 301).

Lord's symmetry property follows simply that the equating function eqy in equations 2.2.1 and 2.2.2 must produce the same results, regardless of which form is used as the starting point—that the inverse of the equating function would produce converted scores on Y that correspond to the observed scores on X. This property rules out consideration for the use of regression techniques as equating procedures since the regression of X on Y is not equivalent to a regression of Y on X.

Available equating procedures rely on either observed test scores, or true scores produced by item response theory (IRT) models. Equating methods include mean-mean (Loyd & Hoover, 1980) which addresses first equity only, and equipercentile, mean-sigma (Marco, 1977), and IRT methods which also address second order equity.

Mean-mean equating methods do not attend to distributional information in the equating function. Either examinee score (observed or true), or IRT item parameter estimates are summarized by their mean values, the differences computed, and those differences used to relate scores on forms X and Y. Mean equating assumes that the distributions of examinee scores are constant, so where this assumption is not met, equipercentile, mean-sigma, or IRT methods may be more appropriate.

Equipercentile procedures equate scores across forms such that the cumulative distribution of the converted scores on the equated form (e.g. new form Y) is equal to the cumulative distribution of the old form (X). Scores on X are converted to percentile ranks and matched with the scores on Y at corresponding ranks to establish the conversions.

Mean-sigma equating includes a consideration of both the score means and distributions of examinee scores on X and Y, enabling the equating functions derived under this method to apply differences across the entire scale. Depending on the assessment purposes, this method can be considered preferable to simple mean-mean equating where application of a constant mean shift, without respect to differences in the distributional properties of examinee scores may results in over- or underestimating ability at different points on the scale. In other words, mean-mean equating results do not satisfy second order equity.

This study uses IRT measurement models due to their pervasive use in testing programs in the US, and because of two desirable characteristics of such models. Item response theory

models separate examinee ability from test difficulty, while reporting both on the same scale, and allowing for equating solutions that are invariant across populations of examinees, at least theoretically. When the two are disentangled, it becomes possible to leverage the item invariant statistical information that is produced for each examinee, and the population invariant statistical information that is produced for items during estimation of the model parameters (Allen and Yen, 1979; Hambleton, Swaminathan, and Rogers, 1991). These IRT parameters, theta (θ) for examinees, and the discrimination (a-parameter), difficulty (b-parameter) and guessing (c-parameter) for items, can be used to produce equating functions.

IRT equating methods are either true score- or observed score-based and differ from each other primarily in how they sum differences in the IRT item parameters (and their corresponding item characteristic curves, ICC). Haebara (1980) summed the squared difference between ICCs for common items, for each examinee within ability levels, whereas Stocking and Lord (1983) summed ICC differences before squaring, making it the squared difference between two TCCs at a given theta, accumulated over examinees. Others (Zeng & Kolen, 1994; Kim & Kolen, 2007) have applied weighted summations over posterior ability distributions. More recently, van der Linden and Barrett (2016) developed the precision-weighted average equating method that differentially weights the contribution of each common item to the equating function based on its IRT parameter estimation error in separate calibrations. The contribution of an item to the equating solution is weighted based on the differences in item parameters between the two calibrations, where items with larger differences (greater IRT estimation error) make smaller contributions to the equating function. To date, this method treats dichotomously scored items only.

Decisions regarding which method is appropriate for a testing program are not necessarily straight forward or easy to make. Many considerations are relevant to decisions about which design and procedures are most appropriate for the testing circumstances, the most fundamental of which is how the scores are intended to be used. For example, a pass/fail decision is less concerned with accounting for score equity across the entire scale, whereas measures of growth along a scale demand are more likely to require that the equating attend to second-order equity.

The administration context is also important including the size and characteristics of the examinee samples, the number of test forms or item pool size required, and the mode of administration e.g. static test forms or adaptive administrations. The administration conditions often constrain design decisions, which ultimately narrows the field of possible equating procedures (Kolen & Brennan, 2014).

Arguably the most common equating design is the NEAT design due to its minimal sampling and administration complexities. It is often preferred over single and random group designs because it defines the equating solution through use of set of common items that can be addressed through test design and construction procedures. Single group designs require longer testing times, risk fatigue effects, and require counterbalanced administrations of two or more forms or blocks of items, and the negative effects of selection bias associated with random groups designs has been demonstrated in multiple studies (Dorans, 1990; Eignor, Stocking, and Cook, 1990).

As mentioned, IRT methods are common in large scale testing programs in the US, primarily due to their ability to separate item difficulty from examinee sample ability and examinee ability from test difficulty. Setting aside mean-mean equating as the least preferred due

to its inattention to likely changes in distributional properties across testing populations, either mean-sigma or characteristic curve methods might be selected for use within a NEAT equating design. Unlike the mean-sigma equating method, however, IRT methods offer a means to account for item discrimination which can offset the influence of larger item difficulty differences where item characteristic curves are otherwise similar (Kolen & Brennan, 2014). For this reason, the current study will focus on the use of both IRT scaling and test equating methods, and are covered in detail in Chapter 3.

Equating results are generally evaluated by how well the equating function maintains equity of scores across forms. In practice, equating results are often evaluated at both the item (for common item designs) and test level (Dorans, Pommerich, & Holland, 2009; Kolen & Brennan, 2014). Items are evaluated for both IRT model fit (e.g. Yen, 1981), and for the presence of any content or statistical differences across administrations. Item difficulty is typically monitored for changes across administrations with simple descriptive statistics (e.g. percent correct or mean as a percent of the maximum possible item score), as well as with standardized measures of differences in the IRT difficulty parameters such as Robust Z (Huynh, 2010). Equating results may also be evaluated at the test level. Depending on the equating design, options here include assessing, 1) how well anchor item IRT parameters correlate across forms, 2) differences in conditional standard errors of the raw to scale score conversion, and 3) evaluation of the standard errors of equating functions (Dorans et al., 2009; Kolen & Brennan, 2014).

Kolen and Brennan (2014) define the standard error of equating as, "…the standard deviation of equated scores over hypothetical replications of an equating procedure in samples from a population or populations of examinees." Carrying out this type of estimation is complex.

Both empirical, such as bootstrap (Efron & Tibshirani, 1993) and analytic procedures (Lord, 1982; Ogasawara, 2000, 2001; Wong, 2015; Barrett and van der Linden, 2017) have been proposed. Bootstrap methods require estimation of the standard error of the parameters used in the equating model by taking multiple random samples with replacement from the test data (Kolen & Brennan, 2014). Analytic methods have the advantage of not requiring such simulations, and rather use asymptotic standard error (ASM) formulas to estimate equating error based on item parameter estimates and the variance/covariance matrix only. One benefit of the use of ASMs over empirical methods such as bootstrap, then, is that equating error for multiple methods can be evaluated and used in the selection of the method associated with the least amount of equating error. Wong (2015), extended the earlier work of Lord (1982) and Ogasawara (2000, 2001a, 2001b) for dichotomously score items to derive ASMs for true score equating under mean-mean, mean-sigma, and concurrent calibration equating procedures with polytomously scored items. Findings suggest that the concurrent calibration, under single group or randomly equivalent group designs, generally produced the lowest equating error, and that error estimates based on the ASM method approximate the empirically estimated error well. The method was not applied to characteristic curve equating methods.

There are many possible sources of error can affect the equating function and threaten the comparability of scores across test forms (Keller & Hambleton, 2013). Potential sources of error include, random equating error due to sampling (Haberman & Dorans, 2009; Kolen & Brennan, 2014; Wong, 2015), IRT parameter estimation error (Sheehan & Mislevy, 1988; Barrett and van der Linden, 2017), violations of the IRT assumptions of unidimensionality (Skaggs and Lissitz, 1986), effects of curricular or test content shifts (Bock, Muraki, & Pfeiffenberger, 1988; Miller & Linn, 1988; Taylor & Lee, 2009), context effects (Yen, 1980; Kingston and Dorans, 1984),

and rater error for polytomously scored items. A few models have been developed to identify and quantify rater bias, including Linacre, (1989), Verhelst and Verstralen (2001); Wilson and Hoskens (2001), Patz, Junker, Johnson, and Mariano (2002), and Casabianca, Junker, & Patz (2016). However, no studies have been identified that directly address the effects of changes in a measure of systematic rater error on test equating solutions, such as those changes that might be expected as automated scoring methods continue to improve.

2.3 Automated Scoring Methods

The first automated scoring tool was developed by Ellis Page (1966) and operationalized in Project Essay Grade (Ajay, Tillett*, &* Page*,* 1973). Page used simple non-linguistic features of essays such as word, sentence, and essay length to predict examinee performance on essay prompts. These features were treated as "proxy" surface features, on which human scores are regressed to develop models to predict essay scores. Although these features continue to be strong predictors of examinee performance, they are also highly criticized for their failure to incorporate other important features of the construct of writing in their models (Perelman, 2014).

A second approach to automated scoring is in the use of latent semantic analysis (LSA) described in Landauer, Foltz, and Laham (1998). This technique is also referred to as a latent semantic indexing or the "bag-of-words" method and uses singular value decomposition (SVD) on a term-by-document matrix to relate frequencies of relevant terms used between examinee responses and master corpuses of text. This approach is less superficial than counting surface features only as it also assesses the content of a response through the term matching process involved in the SVD.

Although these earlier methods showed promise for accurately predicting essay scores, validity demands for the use of more direct features of constructed responses gave rise to the use

of natural language processing (NLP) in automated scoring models. In its simplest definition, natural language processing is method by which written text is automatically parsed into linguistic structures. Manning & Shutze (1999) provide comprehensive coverage of the theory, development, and applications of NLP. For the purposes of this study, the most important aspect of NLP for describing how automated raters use NLP is that the "parses" or combinations thereof are treated as quantifiable features of examinee responses which are in turn used in statistical prediction models, including but not limited to multiple regression, random forest, neural networks, and machine learning techniques. Information extracted from NLP parsers can also be used in rule-based procedures embedded in an AR's overall method. Such procedures can be used to assess direct matches between key words or phrases in the rubric and examinee responses, or such as might be needed to assess topical similarities. This information can be subsequently combined with other features in the prediction models. The number of features used in the model varies dramatically, from a few to thousands (Williamson, Xi, & Breyer, 2012; Bennett & Zhang, 2016).

The training and validation of these statistical models requires a criterion against which the prediction model can be produced. Human rater scores are by and large the most common criterion used in the modeling, where a set of human scores are regressed on to the feature scores and the model validated (and cross-validated) on another set(s) of examinee responses for which human rater scores are available.

Early development of automated raters focused primarily on essay scoring, rising from a need to offer writing students more opportunities for feedback than their teachers had the time to provide (Page, 1966). Project Essay Grade (PEG) was the first program to address this need (Page, 1973), but it was really with the growth of computer-based test administrations in the late

1990s that it became practical to consider operational uses of automated essay scoring. The first automated rater to score essays administered in a large-scale assessment was e-Rater in 1999 when it began scoring essay prompts on the Graduate Management Admissions Test (GMAT®) using a combination of NLP and statistical prediction. Intellimetric®, Intelligent Essay Assessor (IEA), Writing Roadmap™, and a new rendition of PEG, followed in close order with generally similar approaches to essay scoring methods, although IEA is based on NLP and LSA techniques. Like e-Rater, these later automated raters focused on the use of observable "proxy-like" features of student responses to predict essay scores and provide writing feedback. Most automated essay scoring produces analytic and holistic scores that consider the grammar, usage, mechanics, style and organization, and development found in examinee essays. Which features are used and how they are layered or weighted in the statistical prediction models is generally a matter of proprietary knowledge, so can be quite difficult to compare and contrast directly across automated raters.

The primary difference between automated raters for essays appears to be in the features used for prediction and their weights, each seeking to find the combination of observable features that best predicts scores. e-Rater® uses NLP to parse essays into syntactic and grammatical structures and extracts 11 primary features: organization, development, grammar, usage, mechanics, style, average word length, median word frequency, positive features, and two content features. These primary features are based on sub-features that are directly observed in the parse of an examinee's essay (Ramineni & Williamson, 2013). Intellimetric®, on the other hand tags and uses more than 500 linguistic and grammatical features in the score prediction models (Rudner et al., 2006), and Writing Roadmap™ uses approximately 300 features extracted through NLP procedures in its prediction models.

Automated scoring methods for constructed response items share some of the same techniques as essay scoring, but have proven to be a more challenging to develop. Where NLP techniques are very successful at parsing the linguistic features of writing, they are not as directly useful for identifying content related features in examinees responses. Shorter response lengths have also been problematic for the prediction models that require more than less information for accurate predictions. C-rater™, for example, is an automated rater designed to score constructed response items in a variety of content areas. C-rater™ treats the scoring problem as a need to specify the many possible ways a concept might be paraphrased (the "model") and then proceeds to map student responses on to the model via the structures detected in NLP parsing. C-rater™ uses NLP to identify 4 basic structures of a response that include, "syntactic variation" (how the sentence is structured), "pronoun reference" (correct use of pronouns), "morphological variation" (variants of the same word), and the use of similar words. This structure is then used to build a canonical representation that is mapped to the model through a rule-base algorithm (Leacock & Chodrow, 2003).

Although the use of automated raters for essays has grown from its original use on the GMAT, their short-constructed counterparts have not yet found their way into routine operational scoring. In 2013, the second of two automated scoring competitions was held world-wide. The intent of the 2013 competition was to better understand the state of the art of automated scoring for short-text constructed response items and to create an incentive for advancing development and implementation of valid and reliable automated scoring models. The competition was funded by multiple foundations and hosted by a consortia that organized its implementation (Shermis, 2013). The 2013 competition followed an earlier phase in 2012 during which a similar competition focused on automated essay raters. Results of the 2012 competition

largely demonstrated that current automated essay raters score as well or better than humans. The 2013 study involving short-text constructed responses suggested that the methods were not yet sufficient to support the use of automated raters in high stakes assessments (Shermis, 2014).

Liu et al. (2014) used c-Rater™ to score a set of complex science items and evaluated the results. Findings also suggested that agreement rates with human raters were encouraging, but not yet sufficient to replace them for these item types, where two years later, Liu et al. (2016), found that a variant of c-Rater™, c-Rater-ML showed significant improvement in automated to human rater agreement for complex science items scoring. The essential difference is that, rather than attempting to define the domain of possible responses through paraphrases, c-Rater-ML operates much like automated essay scoring with respect to is prediction modeling. This is precisely the type of change in process for automated scoring that is expected to continue, and which might be expected to be problematic for test equating.

Smaller but important improvements in the performance of automated raters is also expected in terms of how they handle various issues that are common to large scale assessment, and those that arise from the very nature of automated scoring. Most testing programs have processes in place that address treatment of unscorable or otherwise flagged examinee responses. Human raters are trained to handle issues such as unreadable text or different language responses and disturbing content. Automated raters, however, need further targeted research to address these elements of student responses that are very easy for humans to process. Complicating these circumstances further, model development and deployment research is finding that examinees learn quite quickly how to change their response processes under automated rater scoring conditions, changing the score distributions between model building and deployment substantially enough to affect the accuracy of the models (V. Kieftenbeld, personal

communication, September 28, 2017). The need for automated rater to achieve greater agreement rates with human raters and to address these specific circumstances strongly suggests that further incremental improvements in scoring accuracy can be expected.

Although the models that employ NLP combined with statistical prediction represent the most frequently adopted in operational assessments, they do come at the cost of a heavy reliance on human scores. That cost has two important implications—financial burden and potential threats to score precision of the scores that are used to train the models. Human scores are expensive to produce, and as will be discussed in section 2.4, they are prone to bias themselves.

In short, although there is some evidence of advancement in short constructed response scoring methods, such as in improved feature development and machine learning techniques, there is significant need for improvement in their scoring accuracy (Williamson et al. 2012; Bennett & Zhang, 2016; Liu et al. 2016; Boyer and Kieftenbeld, 2016). Further, it is increasingly common for test developers to include short constructed response items in test equating designs. The same is not typically true for essay prompts. Consequently, this study will focus on examining the effects of rater bias in scores for short constructed response items.

2.4 Evaluating Rater Quality

2.4.1 Human Rater Quality

Before discussing the details of how automated raters are evaluated, it is important to consider their development in the context of the traditional human scoring processes that precede them. As discussed, most prediction models in automated scoring use human scores to build their models, and are used as the criterion against which rater quality is measured. For most automated rater models in use today, this makes the human rating processes and quality inherently important in establishing and measuring the accuracy of automated raters. Ultimately, problems

with human scoring accuracy and consistency may confound how automated rater quality is established.

Best practices for human scoring processes include attention to training raters, and to evaluation of the scores they produce. Human raters are most often trained on student responses that have been selected by content and scoring experts to represent the range and complexity of the student responses they will be scoring. This provides raters with a strong sense of the kinds of responses they will see during scoring, guidelines for how to apply scoring rubrics in a way that does not reflect their own biases, and some initial practice scoring live responses. This type of training often allows for group discussions of the elements of student responses that make them more or less difficult to score. The goal of this training is to produce raters that accurately and consistently apply the scoring rubrics.

A difficult challenge for human scoring, however, is the tendency toward bias from several possible sources. Rater bias can appear in examinee scores as tendencies to be either more severe or more lenient than expected (Kneeland, 1929; Saal & Landy, 1977; Bernardin, LaShells, Smith, Alvares, 1976). The expectation can be defined as the midpoint of ratings where multiple ratings are gathered for a response (Bernardin, LaShells, Smith, Alvares, 1976; Patz et al., 2002). Another common source of bias is a phenomena called a *halo effect* (Thorndike, 1920) which occurs when high or low performance on one element of the rubric encourages inaccurate scoring of other elements. A third form of rater bias is characterized as raters being reluctant to assign extreme scores. This tendency is referred to as "central tendency" and it can restrict the range of examinee scores (Landy & Trumbo, 1976; McCormick & Tiffin, 1974; Wexley & Yukl, 1977; Zedeck & Blood, 1974).

Leacock, Gonzalez, and Conarro (2014) found that rubrics that are not sufficiently specific may also lead to higher levels of bias as they provide less detailed guidance for scoring examinee responses. This type of challenge can be addressed to some extent by training procedures that include samples and discussion of examinee responses that are particularly difficult to score. Leacock et al. (2014), however, demonstrated large improvements in scoring consistency (for automated raters and humans) where the language used in rubrics was altered in fairly simplistic ways to change its level of specificity.

As human raters conduct scoring, monitoring procedures are also used to track rater consistency. Responses that have been previously scored by expert raters are often seeded into the scoring process and levels agreement between raters and the expert score are computed. Where discrepant agreement is noted, raters can be retrained on the specific issues noted in these types of comparisons. Percent agreement, correlation, kappa (Cohen, 1960) and quadratic weighted kappa (Cohen, 1968) are commonly used for this type of monitoring, as well as for reporting levels of inter-rater reliability in a testing program.

Agreement statistics used for human scoring have played a large, but perhaps insufficient, role in the development and use of automated raters. Where much is known about the sources of human bias in scoring processes, the context in which automated raters are being developed makes it challenging to identify the sources of bias that might be found in automated rater scoring. The sources of automated rater bias are likely method, and even item dependent, but more often than not, the details of those methods are not publicly disclosed.

Beyond rater agreement statistics, other more complex routines might also be used to identify, and importantly, to quantify rater bias, including, the many-facet Rasch measurement model, or Facets (Linacre, 1989), the rater bundle model, (Wilson & Hoskens, 2001), and

hierarchical rater modeling (HRM, Patz et al. 2002) which is described in detail in Chapter 3. In addition to accounting for rater bias in the ratings used by the IRT model to estimate examinee ability, these models can identify raters with anomalous rating patterns, i.e., severity, leniency, and unexpectedly low variability. Such information can be used to identify those candidates who may need additional training, or alternatively, whose scores or services are to be removed from the process.

2.4.2 Automated Rater Quality

Multiple frameworks have been proposed to evaluate the quality of automated raters. From a validity perspective Bennett and Bejar (1998) proposed a system level approach that considers the construct, test design, and task design in the context of the large computer-based systems and validity framework. Expanding the validity framework posed by Bennett and Bejar (1998), Yang, Buckendahl, Juszkieweicz, and Bhola (2002) specified three classes of studies to support the validation of automated scoring systems, studies that assess the 1) relationship between scores produced by different raters, 2) relationship between scores external measures, and 3) scoring processes. The authors conclude that the validation practice for human scores is sufficient to provide validity evidence for automatically produced scores—that automated scores with correspondence to the validated human scores are valid.

In an empirical implementation of the Yang et al. (2002) validity framework, Yang, Buckendahl, Juszkieweicz, and Bhola (2005) suggested statistical criteria for use in the detection of bias in automated raters based on Zwick (1988). Zwick proposed that score distributions between automated raters and human raters first be tested for their marginal homogeneity. Where marginal homogeneity is not rejected, it was proposed that Scott's π coefficient be applied to assess chance corrected agreement (Scott, 1955). The idea underlying this approach was that rater agreement cannot accurately be assessed when rater bias is present.

Williamson, Xi and Breyer (2012) offer a somewhat different framework for the implementation and evaluation of automated scoring systems. Automated scoring is viewed as a continuum of development in which uses are only supported by the technology's ability to meet a set of criteria. High stakes use of automatically generated scores are placed at the most advanced end of the continuum. Specific criteria for determining the adequacy of an automated rater's performance against human rater performance on a single item are posed as 1) a quadratic weighted kappa greater than 0.70, 2) a Pearson product-moment correlation greater than 0.70, 3) a degradation in exact agreement rates from human-to-human and human-to-automated rater less than 0.10, and 4) a standardized mean difference in scores between human and automated raters that less than 0.15.

Most recently, Bennett and Zhang (2016) elaborate a comprehensive approach to establishing the validity of the scores produced by automated raters. The authors acknowledge that attention to all elements of the approach might be quite challenging for most testing programs, but reasons that appropriate evidence of validity reaches far beyond rater agreement and seeks evidence related to score "meaning." Bennett and Zhang point out seven alignments that might serve as sources of validity evidence in a testing program using automated scoring for constructed responses: 1) examinee process and construct, 2) scoring rubric and construct, 3) rater behavior and score rubric, 4) human-to-human ratings, 5) automated raters and aberrant examinee responses, 6) intra-human task-to-task ratings for similar tasks, and 7) human ratings and other indicators. The authors also emphasize the need to assess the invariance of each result across examinee groups.

The most widely implemented type of validation study for automated scoring to date falls in the Yang et al. (2002) first category, the relationship between scores produced by different

raters, and address the fourth source of validity evidence in Bennett and Zhang (2016). In practice this has largely meant a comparison of automated rater scores to human rater scores via descriptive rater agreement statistics, one item at a time.

Leacock & Chodorow (2003) evaluated c-Rater® performance on items from two large scale assessments, the National Assessment for Educational Progress (NAEP) and the Indiana State assessment program. The authors reported item level automated rater to human agreement rates and kappas which ranged from 0.0 to 0.15 degradation from the human to human kappa values.

Rudner, Garcia, & Welch (2006) evaluated Intellimetric® performance on more than 100 essay prompts, reporting rates of automated rater to human agreement and correlations for each. The authors found that Intellimetric® essay scoring performance was generally on par with human raters.

Shermis & Hammer (2013) and Shermis, (2014 & 2015) reported the results of two automated scoring competitons that took place in 2012 and 2013 to assess the state of the art in automated essay scoring and constructed response scoring. These two studies relied primarily on item level quadratic weighted kappas, rank averaged, to compare the performance of many automated raters, finding generally that automated rater essay scores were comparable to human rater scores, and sometimes even outperformed human raters. Findings for the constructed response automated raters were less favorable, concluding that the state of the art was unlikely sufficient for use in operational assessment programs.

Liu, Brew, Blackmore, Gerard, Madhok, and Linn (2014) evaluated the performance of c-Rater® on 4 complex science items, reporting degradations of kappa and mean score differences between human raters and automated raters, as well as Cohens D to assess the

significance of the mean score differences. As mentioned earlier, the authors found that c-Rater® did not perform as well as human raters for these items types. Finally, Liu, Rio, Heilman, Gerard, and Linn (2016) evaluated the performance of c-Rater-ML on 8 complex science tasks, reporting item level means and quadratic weighted kappa values, and finding that c-Rater-ML outperformed c-Rater® on these item types.

A few recent studies have expanded the scope of study of automated (and human!) rater quality. Kieftenbeld & Boyer (2017) proposed an approach based on Demsar (2006) and Garcia and Herrera (2008) for evaluating automated raters over two or more items using inferential methods. Unlike the more typical descriptive methods that are applied one item at a time, the methods investigated in this study use several inferential approaches to examine rater quality over sets of items. The authors concluded that repeated measures ANOVA on ranked quadratic weighed kappas is preferred over more complex non-parametric tests.

An additional aspect of scoring processes that have been more recently discussed in the literature is the concept of "gaming." In the context of the Bennett and Zhang (2016) validity framework, investigations of gaming fall under their fifth source of validity evidence—ensuring that automated raters can appropriately handle aberrant examinee responses. Higgins and Heilman (2014, p. 36) refer to gaming as "construct irrelevant strategies" that examinees use to inflate their scores. The authors propose a framework for approaching this challenge which focuses on the need to anticipate how examinees might plausibly seek to artificially increase their scores, simulating those conditions, and using a proposed gameability metric to evaluate the susceptibility of scoring methods to changes in the metric.

Casabianca, Junker, & Patz (2016), examine automated and human scoring results for a collection of essay prompts using the HRM (Patz et al. 2002). This study found that the

automated rater performed was slightly more severe than human raters, but at a level determined to be not statistically significant. Some details of the model are included here to demonstrate their potential usefulness in the field of automated scoring, and specifically, the relevance of HRM to this study.

In the application of measurement models, the HRM accounts for the dependence of multiple ratings within items to estimate and distinguish levels of rater bias and variability in the essay scores for each rater. The model estimates *ideal* ratings that account for this dependence among ratings and can then be used in IRT scaling models. The rater bias and variability estimates produced, however, might be further useful in the evaluation of automated raters. Disentangling the bias present in human scores and providing estimates of the *ideal* ratings produced in the HRM (which are merely assumed in IRT models) are two potential advantages.

First, illuminating the level of rater bias that is in the human scores that are used to train and validate automated raters might allow the opportunity for model training and validation effort to leverage this specific information about rater bias to select the "best" scores on which to train the model. Second, and specific to this study, measures of rater bias and variability might be used to characterize improvements in the field of automated rater scoring based on increments of bias reduction. This may be particularly true in the current context of automated rater development, where the sources of rater bias are not well known and may not generalize well over methods and items.

2.5 Psychometric Implications for Test Equating

As automated scoring methods continue to advance and produce scores that are increasingly consistent with human raters, reductions in the amount of automated rater bias in the scores for constructed response items might be expected. Although this reduction in error is

desirable, one implication of such reductions is that they represent changes in the amount error that is propagated in the estimates for IRT item parameters, examinee thetas, and consequently, in the equating functions. Where constructed response items are desired in equated tests, there is a need to evaluate the stability of equating functions across these changing levels of rater bias.

The IRT estimation methods used most often in operational settings assume that rater error is not present in the scores used to produce the measurement scales, and essentially ignore this possible threat to score comparability over changes in automated rater quality (Casabianca et al., 2016). Possible scenarios where it will be important to understand the effects of changing rater bias include situations where different automated raters are used to score the same items over time, where improvements are made to automated raters that score the same items over time, and where examinee behavior changes over time in reaction the use of automated raters.

Although many studies have examined the effects of sampling and IRT estimation error, no studies have been found that examine the impact of rater bias on test equating functions. Typical equating designs and procedures under which automated rater scores might conceivably be used have been described in this chapter, along with various types of automated raters and the methods used to assess the validity and consistency of their scores. These descriptions demonstrate the context in which score comparability might be threatened due to changes in rater bias inherent to improvements in the state of the art. The specific methods that will be used in this study to investigate the effects of changes in rater bias on test equating functions draw from of those described, and are elaborated in detail in Chapter 3.

CHAPTER III

METHOD

3.1 Overview

This study begins with the assertion that automated scoring methods for constructed response items will continue to improve over time, and that their current and future changes (presumably reductions) in levels of systematic rater error may impact the comparability of scores across those rater improvements. As the fundamental purpose of equating is to be able to make generalizations from one observed score to other potentially observable scores on alternate forms based on the same test blueprints, the circumstance of changing levels of rater bias and variability may threaten such generalizations. Further, and since there any number of procedures that may be used to equate alternate forms, it will be useful to identify a means to compare impact across equating methods.

An immediate challenge for answering this study's research questions is the identification of an appropriate definition of systematic rater error. Interrater agreement statistics, such as percent of agreement (perfect, adjacent, and non-adjacent discrepancies), kappa, and weighted kappa are commonplace due to their practicality, but they largely assume that there is a set of human produced scores that can be accepted as sufficiently bias free for use in training and validation of the models. This presents a fundamental limitation in current operational practice for evaluating rater quality due to the confounding that is inherent in using agreement with human raters alone as the primary means to assess score quality. There are few methods available that provide statistical measures of the level of systematic rater error (Linacre, 1968; Verhelst & Verstralen's, 2001; Wilson & Hoskens, 2001; Patz et al., 2002). Based on a current

review of published studies, only one study is found (Casabianca et al, 2016) to have applied such a technique to assess automated rater quality.

The Facets model (Linacre, 1968) models rater effects as interactions between examinee responses, items, and raters, however, it fails to account for the dependence of ratings within items for a given examinee ability (Patz et al, 2002). Due to the limitations of simple rater agreement statistics and to the Facets model, then, an alternative measure is needed. The HRM model (Patz et al, 2002) provides a solution to this problem through use of two measurement stages, where the first stage functions as a signal detection model to establish ideal ratings for use in the second stage production of IRT parameter estimates. Other models have also been proposed that account for the dependence of ratings within items for a given examinee ability, such as the rater bundle model (Wilson & Hoskens, 2001), and the Verhelst and Verstralen's (2001) IRT based model.

The HRM is used in this study to introduce noise into examinee scores simulated based on IRT models. The HRM was chosen due to the relatively straight forward way that the signal detection component can be used to specify different levels of bias and variability into examinee scores. Following the data simulation and introduction of defined levels of bias and variability, the tests were placed on IRT scales and equated. Results were evaluated by how well a regression of alternate form (F2) raw scores on equated F2 raw scores predicts *ideal* (noise-free) raw scores.

3.2 Data Simulation

The data for this study were simulated based on two test designs. Both designs include two 60 item, unidimensional, mixed-format tests. The two designs differ only in their concentration of constructed response items. One design includes 5% constructed response items

and the second includes 10% constructed response items. To assess the impact of a variety of constructed response item types with 3, 4, and 5 score levels (maximum points equal to 2, 3, and 4) were included in the test designs. Although essay items are typically worth 5 or 6 maximum points, it would be not be typical to include such items in an equating, thus the selection of constructed response item sets worth less than 5 points maximum. One of each constructed response item type was included in the 5% constructed response test design. This was doubled for the 10% test design to assess the impact of increasing numbers of constructed response items in the equating. All other items were simulated as dichotomously scored multiple-choice items. Although approximately 2000 examinees would be sufficient for estimating IRT parameters under the models specified below, 10000 examinees were simulated for each test form to minimize possible random error in the equating functions. Table 1 provides a summary of the test designs simulated in this study.

Table 1 Test Design and IRT Simulation Parameters, Two Equating Design

Test Design	% CR	No. CR Items	No. CR Items	No. CR Items
		1	1	1
		1	1	1
1	5	1	1	1
		1	1	1
		1	1	1
		2	2	2
		2	2	2
2	10	2	2	2
		2	2	2
		2	2	2

CR=Constructed Response

The IRT models used to simulate the data are the 3-parameter logistic model (3PL, Birnbaum, 1968) for dichotomously scored items, and the Generalized Partial Credit model (GPCM, Muraki, 1992) for polytomously scored items. The 3PL model produces the probability of a correct response to a dichotomously score item given examinee ability, and is shown,

$$P(x_i = 1|\theta_p, a_i, b_i, c_i = \ c_i + (1 - c_i)\frac{\exp[Da_i(\theta_p - b_i)]}{1+\exp[Da_i(\theta_p - b_i)]} \quad (1)$$

where $P(x_i = 1|\theta_p$ is the conditional probability of a correct response for examinee p on item i given θ_p, and θ_p is examinee ability, b_i is location parameter (item difficulty), a_i is the slope parameter (item discrimination), c_i is the height of the lower asymptote or examinee "guessing" where $c_i > 0$, and where D is a scaling constant set to 1.7 to approximate the normal ogive.

The GCPM model produces the probability of choosing response level k over $k - 1$ for an examinee at a given ability level, given by,

$$P(x_i = 1|\theta_p, a_i, b_i, d_{ia} = \ \frac{\exp[\Sigma_{v=0}^{k} Da_i(\theta - b_i - d_{ia})]}{\Sigma_{c=0}^{A_i} \exp[\Sigma_{v=o}^{c} Da_i(\theta - b_i - d_{ia})]} \quad (2)$$

where d_{ia} is the category parameter for a given item score point.

The IRT models in equations (1) and (2) assume that the item responses *x* are observed without error, but imprecision in the rating process for constructed-response items implies that this assumption is not true for these items. The HRM was used to simulate a range of realistic noise in the rating process, but before describing the scenarios and this particular use of the HRM in detail, it is useful to provide some detail on the model as presented in Patz et al. (2002), and Casabianca et al. (2016). The HRM is a three-level hierarchy where observed ratings (X_{ipr}) are nested within an ideal rating (ξ_{pi}), within examinee true scores (θ) as,

$$\begin{cases} \theta_p \sim i.i.d.N(\mu, \sigma^2), \quad p = 1, \dots P \\ \xi_{pi} \sim \text{an IRT model}, i = 1, \dots, I, \text{for each } p \\ X_{ipr} \sim \text{ signal detection model}, r = 1, \dots, R, \text{for each } p, i \end{cases} \quad (3)$$

for completely crossed designs. Incomplete designs treat missing data as missing completely at random (MCAR).

The signal detection model uses a discrete unimodal distribution for each row of a matrix of response probabilities which define the relationship between X_{ipr} and ξ_{pi}. The mode of this distribution is the rater bias (ϕ), and the variability represents the unreliability of a rater (ψ). Refer to Table 2, reproduced from Patz et al. (2002) to illustrate the signal detection process.

Table 2. The Matrix of Rating Probabilities Describing the Signal Detection Process Modeled in the HRM

	Observed Rating (k)				
Ideal Rating (ξ)	0	1	2	3	4
0	P_{00r}	P_{01r}	P_{02r}	P_{03r}	P_{04r}
1	P_{10r}	P_{11r}	P_{12r}	P_{13r}	P_{14r}
2	P_{20r}	P_{21r}	P_{22r}	P_{23r}	P_{24r}
3	P_{30r}	P_{31r}	P_{23r}	P_{33r}	P_{34r}
4	P_{40r}	P_{41r}	P_{24r}	P_{43r}	P_{44r}

*Reproduced from Patz et al. (2002)
Note. $P_{\xi kr}$ ≡ P[Rater *r* rates *k* | ideal rating ξ in each row of this matrix

Within each row of the matrix probabilities are assumed to follow a normal distribution with mean $\xi + \phi$, and standard deviation ψ,

$$p_{\xi ar} = P\left(X_{pir} = a \middle| \xi_{pi} = \xi\right) \propto \exp\{-\tfrac{1}{2\psi^2}[a-(\xi + \phi_r)]\} \qquad (4)$$

Consequently, when ϕ=0, a rater is likely to score consistently with ideal ratings. When ϕ is negative, the rater is more likely to rate examinee responses more severely than when it is 0, and when ϕ is positive the rater is more likely to be more lenient in their scoring. When ψ is 0 a rater is considered to be perfectly consistent. When both bias and variability parameters are zero, the matrix in Table 2 is a unit diagonal matrix, meaning the observed rating matches the ideal rating with probability 1. The HRM is most often estimated using Bayesian techniques, where prior distributions are specified for IRT and signal detection parameters, but estimation procedures are not be elaborated here as the model was leveraged to introduce systematic error into simulations of ideal ratings.

The bias scenarios examined here include ideal, human, and automated rater scenarios so that the hypothesized automated rater scenarios might be better understood relative to more typical human rater scenarios. As this study relies on a simulation of these scenarios, a fundamental task was to define the difference, in terms of rater bias and variability, between humans and automated raters. Casabianca et al (2016) provides information that can serve as a basis for these definitions in terms of how the HRM characterizes noise in examinee scores due to systematic rater error. Study results specified average human and automated rater bias and variability estimates for a set of examinee scores on a 6-level essay prompt. These values were used as a starting point to define bias and variability for both rater types, but for the automated raters, were then adjusted proportionally for each item type included in the study, i.e. 5-, 4-, and 3-level constructed response items.

An important consideration in the data simulation process was to identify a correspondence between the bias and variability values used here and current operational practices for automated scoring model selection and use. Although other criteria for evaluating rater quality can be used in practice, the use of a 0.70 quadratic weighted kappa is often viewed as a minimum threshold for accepting the sufficiency of an automated rater for operational uses (Williamson et al, 2012). So, although the rater bias values used in this study are taken from the Casabianca et al (2016) results directly (with the noted proportional reductions), the variability (or unreliability values) were assigned such that the highest level of automated rater bias and variability corresponded to about a 0.70 quadratic weighted kappa.

The IRT parameter distributions used to simulate the ideal ("before noise") data were selected with the goal of producing typically reliable and reasonable quality test scores. The examinee theta distributions and item difficulties for Form 1 are assumed to be normally

distributed $N(0,1)$ so that both ability and item difficulty are similarly centered and distributed about the measurement scale when establishing the base scales of measurement. Form 2 was simulated with a small change in difficulty, where b is N(0.1,1). This choice is somewhat arbitrary but is intended to mimic typical empirical scenarios where the item difficulty distribution is targeted during test construction to align with the theta distribution and where overall form difficulty varies (i.e. the reason we equate). The IRT item parameter distributions used for a and c are Log-normal(0,0.2) and $U(0, 0.3)$ respectively to approximate a reasonably well discriminating test with lower levels of guessing on dichotomously scored items. These IRT parameter distributions were used for all simulation scenarios.

Although one purpose of the NEAT is to equate tests with non-equivalent groups, the data simulated for use in a NEAT equating here assumed the same distribution for both forms to avoid any confounding of the results related to examinee ability. Simulations for the single group design used the *exact same* theta distribution for each scenario. The details of the equating design and procedures are discussed in the analysis sections below. The IRT simulated data with ideal ratings were produced using WinGen (Han, 2007) and the HRM-based noise was introduced using R version 3.1.3 (R Core Team, 2013).

The first step in the simulations then, was to produce examinee score arrays for test designs 1 and 2 based on the defined IRT parameter distributions, bias, variability, and item score ranges. Scenarios with ideal ratings are assumed to be free of rater bias and variability as expected by the IRT models used to produce them, so are essentially equivalent to application of HRM where $\phi = 0$ and $\psi = 0$.

As the Casabianca et al (2016) study concluded that scores from the operational automated rater were not statistically significantly different from the human raters, the current

study starts from an assumption that the levels of rater noise detected in Casabianca et al. are acceptable for use in operational settings. As the items examined in Casabianca et al. contained 6 levels, the bias values were reduced proportionally for the 5-, 4-, and 3-level items in this study. Further, as one current objective is to evaluate the effect of improvements in automated rater scoring, the values for ϕ and ψ were improved in from form 1 to form 2, and compared to results with simulated human and zero noise scenarios, in addition to a scenario where the automated rater noise is held constant over forms 1 and 2. The ideal scenario is used as a baseline for comparison of equating results with each rater noise scenario. In this way, each rater noise scenario may be assessed with respect to a theoretically ideal set of scores. The human scenario is included to provide a point of comparison between automated and human raters, and the automated rater scenario with constant rater bias and variability is included to provide a point of comparison to determine the equating impact of improvements in the state of the art in automated scores.

Finally, to relate findings to current practice for accepting minimum sufficiency for the use of automated raters, variability for the automated rater scenario was assigned to each item type based on the proportionally reduced bias and correspondence with approximately 0.70 quadratic weighted kappa (QWK). For humans, variability was reduced slightly for each drop in max score levels, by 0.1, attempting to keep human variability at or above automated raters to mimic empirical expectations. The bias and variability values and their corresponding QWKs are summarized in Table 3.

Table 3 Bias and Variability Scenarios

Simulation Scenario	No. CR Levels	Form 1 QWK	Form 1 Bias	Form 1 Variability	Form 2 Bias	Form 2 Variability
Automated Rater (Constant Noise)	3	0.74	-0.116	0.50	-0.116	0.50
	4	0.70	-0.155	0.50	-0.155	0.50
	5	0.71	-0.194	0.80	-0.194	0.80
Automated Rater (Reduced Noise)	3	0.74	-0.116	0.50	-0.058	0.25
	4	0.70	-0.155	0.50	-0.078	0.25
	5	0.71	-0.194	0.80	-0.097	0.40
Human (Constant Noise)	3	0.58	-0.002	0.50	-0.002	0.50
	4	0.61	-0.003	0.60	-0.003	0.60
	5	0.75	-0.004	0.70	-0.004	0.70
Ideal (No Noise)	3	1.00	0.000	0.00	0.000	0.00
	4	1.00	0.000	0.00	0.000	0.00
	5	1.00	0.000	0.00	0.000	0.00

CR=Constructed Response

3.3 Analyses

3.3.1 Single Group Equating

For the single group equating analyses, all items across test forms 1 and 2 were treated as different items. The absence of fatigue and order effects is assumed. In this case, Forms 1 and 2 were concurrently calibrated, which was designed to mimic a single group of examinees taking both forms. The quality of the single group equating solutions were evaluated through an analysis of the ideal raw score prediction rates of the equating solutions. Mean IRT item parameters for constructed response items are also provide across bias scenarios to show the effects of bias on these parameters.

3.3.2. NEAT Equating

Using the IRT models noted in equations 1 and 2, the base measurement scales were set on Form 1 for each rater type (automated, human, and ideal). The IRT parameter estimation software, PARSCALE (Muraki & Bock, 2003) was used to estimate thetas and item parameters for each scenario.

To implement the NEAT design, the first step following establishment of the base measurement scale was to select a set of 40% of the items to function as common (or "anchor") items across test forms. Forty percent was chosen to provide fairly robust anchor size. These common item sets, were selected randomly to approximate the statistical characteristics of the total test. However, all constructed response items were retained in the anchor as the purpose of this study is to investigate the effect of systematic constructed response scoring error on test equating results. As an aside, it is noted that for any common constructed response items not included in the anchor set under a NEAT design, new item parameters items typically estimated, so the impact of changes in scoring quality for these items would be reflected in the new item parameters and would not pose a threat to the equating.

To leverage the features of scale independence from examinee ability in IRT scales, the tests under the NEAT design were equated using IRT equating. Specifically, Stocking and Lord (1983) linear transformations which are designed to minimize the average squared differences between true-score (θ) estimates for examinee groups were used. After separate form calibrations, the linear transformation that minimized the average squared difference between anchor item ICCs across forms was determined, where the minimization is defined by F which is a function of transformation constants M_1 and M_2,

$$F = \frac{1}{N}\sum_{j=1}^{N}(\hat{\gamma}_j - \hat{\gamma}_j^*)^2 \tag{5}$$

where N is the number of examinees in a group, $\hat{\gamma}_j$ is the estimated true score obtained from the base test form, and $\hat{\gamma}_j^*$ is the estimated true score obtained from the equated test form after it has been transformed to the previous scale as follows:

$$\hat{\gamma}_j = \hat{\gamma}(\theta_j) \sum_{i=1}^{N} P_i(\theta_j ; a_i, b_i, c_i) \tag{6}$$

$$\hat{\gamma}_j^* = \hat{\gamma}(\theta_j) \sum_{i=1}^{N} P_i(\theta_j ; \frac{a_i}{M_1}, M_1 b_i + M_2, c_i) \tag{7}$$

where a_i, b_i, c_i are the IRT discrimination, location (difficulty), guessing parameters for item *i* from equation (1). Equating was performed using IRTEQ (Han, 2009).

3.3.3. Examination of Equating in a Prediction Framework Using Data Replications

As discussed in Section 2.2, the fundamental problem of equating is to provide scores for different test forms, or testing instances that are sufficiently comparable to allow for the same inferences to be made about each score on the scale, regardless of which form or set of items an examinee sees. Examinee scores for equated forms are adjusted through some equating procedure to put them on the scale of the baseline form, and the evaluation of equating quality may be focused on the accuracy with which the equating procedure produces scores that agree with observed scores on the baseline form.

Without loss of generality the convention of having "form 1" indicate the baseline form is adopted, and "form 2" indicates a different test form that is to be made comparable to form 1 through an equating procedure. In the case of single group equating forms 1 and 2 have no items in common, whereas forms 1 and 2 studied in NEAT designs do have items in common as described in Section 3.2. The term "form 1 score" denotes the total number of raw score points earned by an examinee on form 1, and "form 2 score" has the same meaning for form 2. Note that in general "form 1 score" could be some other function of the raw item vector such as an item pattern score. The term "equated form 2 score" denotes the result of applying an equating procedure to a form 2 score to make it directly comparable to form 1 scores (i.e., to "place it on the form 1 scale"). Equated form 2 scores may thus also be considered "predicted form 1 scores,"

which is useful in the context of this study because actual form 1 scores are also present and methods for quantifying prediction accuracy are directly applicable as measures of equating quality.

It is noteworthy that studies of prediction accuracy are common in the development and validation of automated raters, where the focus is on the fidelity with which the algorithms in the automated raters are able to predict human scores. In these and related contexts the complete data on which the algorithm is built (i.e., "training data") is segmented from the data on which the accuracy of the algorithm is evaluated (i.e., "validation data"), since this allows evaluation of the accuracy of predictions that is not spuriously inflated by model over-fit. A similar approach is adopted in our simulation studies: one set of data is used to derive the equating functions and a replicated data set is used to evaluate the quality of the equating in terms of the accuracy of the cross-form predictions it produces.

In particular, the primary statistics used to evaluate the equating quality will be Pearson correlation and root mean squared error:

$$Cor(x,\hat{y}) = \frac{n(\Sigma x\hat{y}) - (\Sigma x)(\Sigma \hat{y})}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2][\,\Sigma \hat{y}^2 - (\Sigma \hat{y})^2]}}$$

Where x is the form 1 score, and $\hat{y}$ is the equated form 2 score (aka "predicted form 1 score"), and

$$RMSE(x,\hat{y}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - x_n)^2}$$

where x_n is the form 1 score for examinee n, and $\hat{y}_n$ is the equated form 2 score (aka "predicted form 1 score").

In addition to these primary raw score prediction-based measures equating quality, we will also examine results from three other perspectives: 1) using the underlying latent variable available in our simulation study context, and 2) using indicators typically employed to evaluate equating based on real data collected in equating study designs. For the NEAT equating this includes an examination of the means for both the discrimination and difficulty item parameters, the correlation between item parameters for the common items, and the minimization function, F,

The replication data for the NEAT equating scenario was produced in a slightly different manner than for the single group. Since the data on which the NEAT equating is performed represents different groups taking different forms with common items, an approach that capitalizes on the fact that simulated examinees can be administered the same items without effect is needed. In this case, the replicate response data used for "validation" was simulated using the originally simulated thetas for form 2, and the item parameters used to simulate the original response data for forms 1 and 2. Although this type of forms administration would not typically occur in operational settings due to memory effects for common items, no such problem exists in the simulation. Further, the simulation of this data allows for an assessment of the impact of different rater noise levels that can be compared across the single group and NEAT equating scenarios.

3.3.4. Examination of Equating Impact on Examinees

To further examine differences in equating quality across rater scenarios and test designs, examinee theta estimates are compared. Recall that the data in this study are simulated to include constructed response scores for groups of examinees that are theoretically scored by ideal, typical human, and automated raters, for two different test designs. The only thing that varies across the scenarios is the amount of rater noise introduced and the total number of constructed response items. This means that, after the 2 forms in each scenario are equated, the theta

estimates based on equated test forms can be compared across scenarios in a way that isolates the impact of the specified changes in rater bias and variability.

Specifically, form 2 theta estimates, based on the human rater scenario are compared to form 2 theta estimates based on ideal ratings, for both test designs. Then the form 2 theta estimates, based on the automated rater scenario are compared to the same form 2 theta estimates based on ideal ratings, for both test designs. The comparisons are performed in a similar manner to the preceding raw score analyses. In this case, the correlation and RMSE between the pairs of noisy and ideal theta estimates are examined for changes across the scenarios, but a direct comparison of the theta estimates is made sequentially between each noisy rater scenario and its corresponding ideal scenario.

Last, a single cut score is set in the center of the simulated form 1 ability distribution (i.e. $\theta=0$), to examine examinee classification impact between the human and ideal results, and the automated rater and ideal results, across the two test designs, and equating procedures. Setting the cut score at 0 is intended to assess impact where the most examinees are located.

CHAPTER IV

RESULTS

4.1 Summary of Impact to IRT Item Parameters

Before examining the equating results, the constructed response item parameters means, (before equating) were reviewed to understand item parameter changes under the various bias and variability scenarios. An important result to be considered throughout each finding discussed here is that the concurrent IRT calibrations under the single group equating design did not result in parameter estimates for 2 of the 3 constructed response items in test design 1 (for all scenarios), and for 2 of the 6 items in test design 2 (automated rater, constant noise scenario). Consequently, these items were dropped from the remainder of the single group equating analyses, which negatively influences an ability to compare results across equating designs, and across all rater scenarios with the single group design.

Overall, item difficulty may be influenced by the levels of rater bias and variability applied in this study, but this influence does not appear to be systematic. Since the human and automated rater scenarios are different from the ideal only in that the defined levels of rater bias and variability have been introduced into the examinee scores used for calibration, the ideal is used as the point of reference for what the mean IRT item parameters would be without rater bias and variability. In this way, the ideal is a benchmark (although based on an indeterminate IRT model, the GPCM) against which changes in calibration and equating outcomes might be measured. Looking at the mean constructed response IRT parameter estimates resulting from the single group concurrent calibration, then, the differences between the human, automated, and ideal raters scenarios appears to be inconsistent. The largest differences are noted for the automated rater scenario where the noise levels across forms are constant at the higher levels.

For the NEAT results there is a clear difference between ideal and rater noise scenarios for design 1 form 1, where difficulty decreases by about 0.2 from the ideal to both noise scenarios. On the contrary, for design 2, form 1, there is a clear increase in difficulty both automated rater scenarios. Last, there is large increase in difficulty of the constructed response items for the automated rater (constant noise) scenario from the ideal scenario. All other design and form comparisons show similar levels of difficulty across designs, forms and rater noise scenarios. These results suggest that the impact on IRT difficulty of the bias and variability levels presented in Table 3 may not be predictable.

There is, however, a notable and systematic decline in item discrimination between ideal and noisy data, as might be expected. The difference between the human and automated rater scenarios is small except for Form 2 under both test designs, for the automated rater (reduced noise) scenario. In the case of the reduced noise scenario, form 2 has half the bias and variability of form 1. These results indicate that rater variability may be quite influential on item discriminations at these relatively low levels of bias (i.e. <.5). Tables 4 and 5 show the IRT parameter means for the single group and NEAT calibrations. The IRT item parameters for form 1 and post-equated form 2, for all designs and rater scenarios, are provided in the Appendix for reference.

Table 4 Mean Discrimination and Difficulty of Constructed Response Items, Single Group

	SG Concurrent Calibration IRT Item Parameter CR Linking Item Means by Design and Form							
	a-Parameter				b-Parameter			
Rater Scenario	D1 F1	D1 F2	D2 F1	D2 F2	D1 F1	D1 F2	D2 F1	D2F2
Ideal	1.11	0.88	0.95	0.97	-1.78	0.08	-0.06	0.23
Human	0.44	0.55	0.57	0.55	-2.23	0.04	-0.02	0.27
Automated (Constant Noise)	0.39	0.56	0.57	0.55	-1.10	-2.01	0.31	0.31
Automated (Reduced Noise)	0.40	0.80	0.57	0.55	-2.14	0.10	0.14	0.26

D=Design; F=Form; CR=Constructed Response

*Note that for design 1, 2 of the 3 constructed response IRT item parameters could not be estimated, so only 1 3-level constructed response item is represented in this summary and in all equating analyses. For design 2, only 2of the 6 items were retained in the analyses for the automated rater (constant noise) scenario.

Table 5 Mean Discrimination and Difficulty of Constructed Response Items, NEAT

	NEAT (Separate Calibration Before S&L Equating) IRT Item Parameter CR Linking Item Means by Design and Form							
	a-Parameter				b-Parameter			
Rater Scenario	D1 F1	D1 F2	D2 F1	D2 F2	D1 F1	D1 F2	D2 F1	D2F2
Ideal	0.940	0.981	0.956	0.962	-0.723	-0.560	0.013	-0.005
Human	0.522	0.560	0.588	0.591	-0.550	-0.560	0.003	-0.011
Automated (Constant Noise)	0.522	0.571	0.570	0.466	-0.560	0.169	0.169	0.051
Automated (Reduced Noise)	0.522	0.810	0.570	0.800	-0.560	-0.522	0.169	0.057

D=Design; F=Form; CR=Constructed Response

4.2 Raw Score Impact

Tables 6 and 7 show the RMSE and correlation results for a comparison of x and $\hat{y}$, where x is the total observed raw score on form 1 and $\hat{y}$ is the total raw score predicted by the equating solution. The RMSE values range from 4.23 to 5.00 for the single group equating results and from 4.70 to 4.95 for the NEAT results. The RMSE values for the human rater scenario and automated rater scenarios are very similar within test designs, with an overall pattern that appears to very slightly favor a reduction in rater bias and variability across the equated forms. For design 1, an RMSE of 4.25 for the reduced noise scenario is 0.01 better than humans, and 0.05 better than the automated rater scenario based on the application of the uniformly higher levels of bias and variability across the test forms. A result that is not consistent with the expected pattern occurs in the single group analyses for design 2 where the RMSE is 4.41 for automated rater (reduced noise) and 4.76 for the ideal rater scenario. The expected pattern would clearly be for a larger RMSE to be observed where rater bias and variability are present than where it is not. This suggests a possibility that the RMSE values resulting from these analyses may be within random variation. This unexpected result does not occur in the NEAT equating results where both the design 1 and design 2 RMSE values have notably smaller variations, but higher, than for the single group results.

The variability in correlations between x and $\hat{y}$ is very small, ranging from 0.93-0.95 for single group, and 0.94-0.95 for NEAT. In both cases, the correlations suggest a very strong relationship between observed and equated raw scores as would be the target of any equating procedure. The degradation in correlation between the ideal and noise scenarios is either extremely small, or non-existent, suggesting that the bias and variability values used here do not have a meaningful influence on the correlation between observed and predicted raw scores.

Table 6 Summary of RMSE and Correlation between x and $\hat{y}$, Single Group

	RMSE		r	
Rater Scenario	Design 1	Design 2	Design 1	Design 2
Ideal	4.23	4.76	0.93	0.95
Human	4.26	5.00	0.93	0.94
Automated (Constant Noise)	4.30	4.70	0.93	0.93
Automated (Reduced Noise)	4.25	4.41	0.93	0.94

Table 7 Summary of RMSE and Correlation between x and $\hat{y}$, NEAT

	RMSE		r	
Rater Scenario	Design 1	Design 2	Design 1	Design 2
Ideal	4.70	4.73	0.94	0.95
Human	4.85	4.95	0.94	0.94
Automated (Constant Noise)	4.85	4.81	0.94	0.95
Automated (Reduced Noise)	4.78	4.81	0.94	0.95

Figures 1 through 16 show the relationship between the equated raw scores and observed form 1 scores along the full raw score scale. This relationship shows a barely perceptible tendency for rater noise to bias scores downwards at the upper end of the scale and upwards at the lower end. Figures 3, 4, 7 and 8, which shows the raw score relationship under test designs 1 and 2, for both automated rater scenarios illustrates the bias effect at the lower end of the raw score scale most clearly, and the effects are slightly stronger under the single group equating scenario than under the NEAT. In test design 2, the automated rater with reduced noise levels across the forms shows a clear equating bias at the upper end of the raw score scale (Figure 8).

As might be expected based on the RMSE patterns in Tables 6 and 7, the human rater scenario patterns are similar the automated rater patterns for both the single group and NEAT results.

Although these effects appear to be small, they are systematic across test and equating designs, and across rater scenarios. Importantly, the effect is consistently more notable for test design 2 where there are 6 versus 3 constructed response items in the test design, noting that the single group design is not directly comparable across equating scenarios as only one constructed response item was used for equating. For design 2, it does appear that the human rater scenario produces a more noticeable effect at the lower end of the scale and that the automated rater produces a more noticeable effect at the upper end of the scale. Refer to Figures 1-16 to see displays of each rater scenario, under each test and equating design.

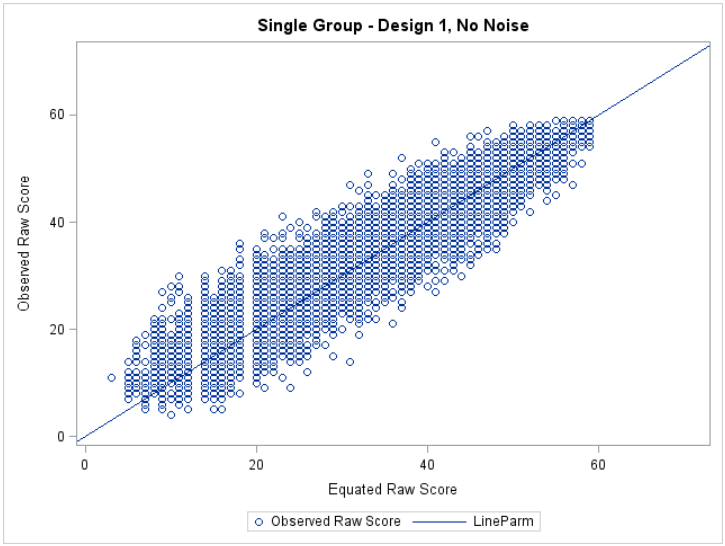

Figure 1 Plot of x and $\hat{y}$, Single Group D1 Ideal

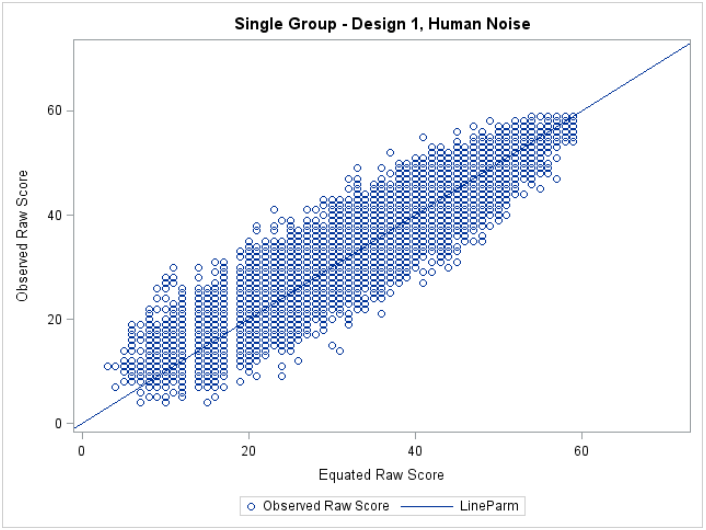

Figure 2 Plot of x and $\hat{y}$, Single Group D1 Human

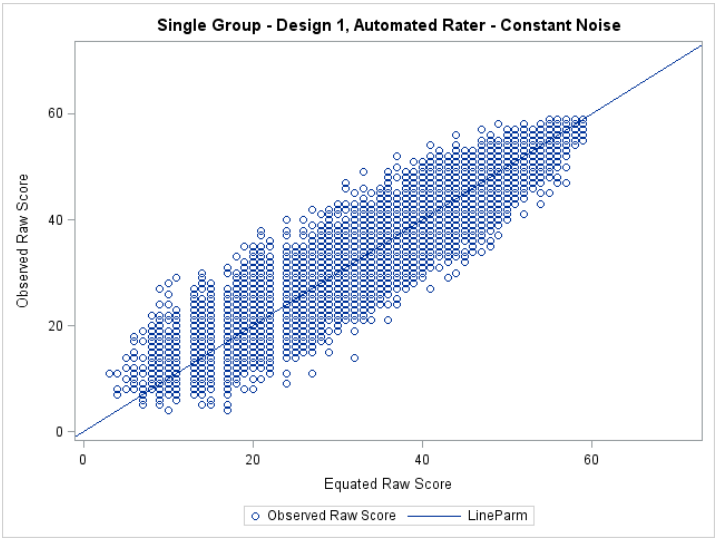

Figure 3 Plot of x and $\hat{y}$, Single Group D1 Automated Rater (Constant Noise)

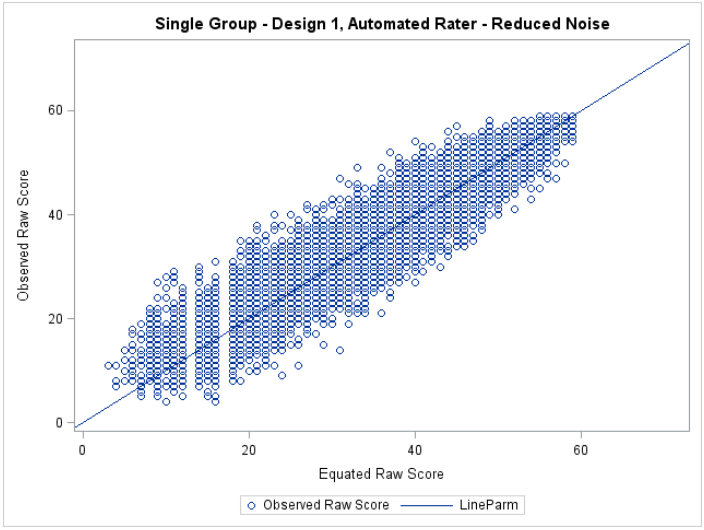

Figure 4 Plot of x and $\hat{y}$, Single Group D1 Automated Rater (Reduced Noise)

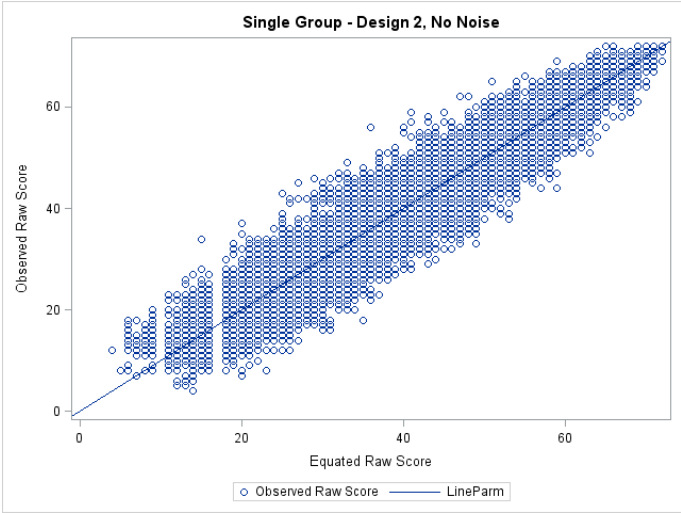

Figure 5 Plot of x and $\hat{y}$, Single Group D2 Ideal

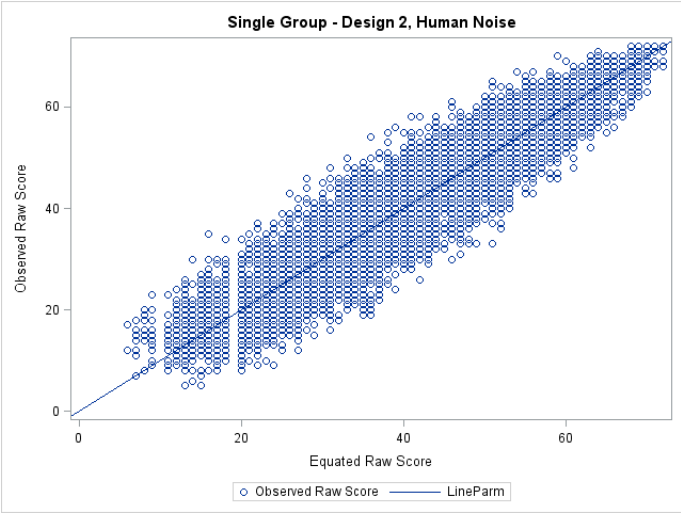

Figure 6 Plot of x and $\hat{y}$, Single Group D2 Human

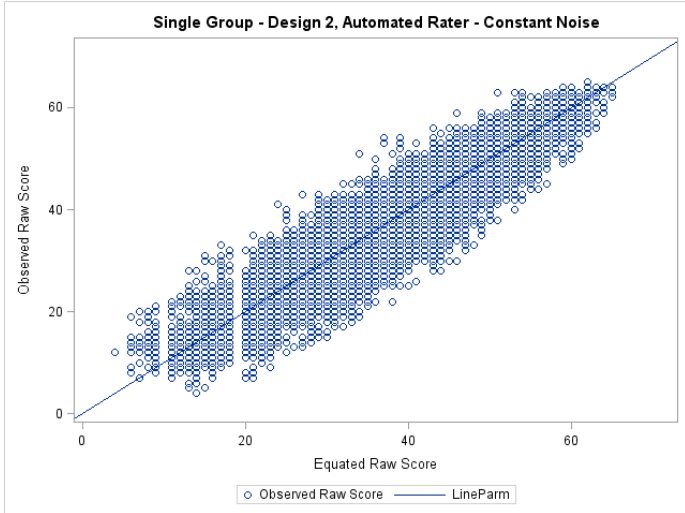

Figure 7 Plot of x and $\hat{y}$, Single Group D2 Automated Rater (Constant Noise)

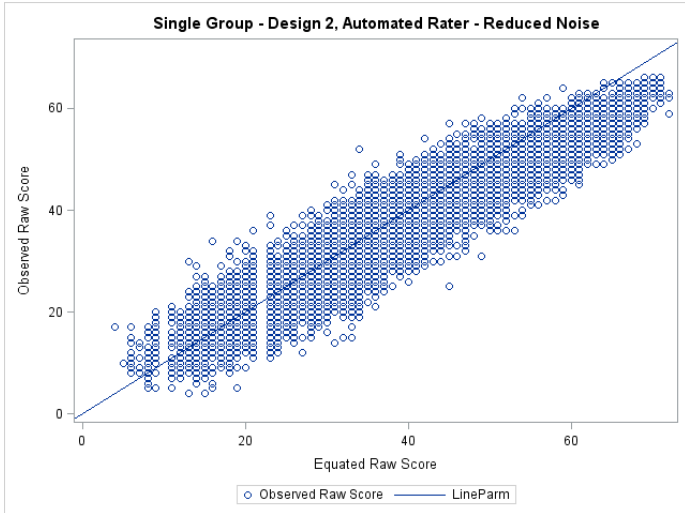

Figure 8 Plot of x and $\hat{y}$, Single Group D2 Automated Rater (Reduced Noise)

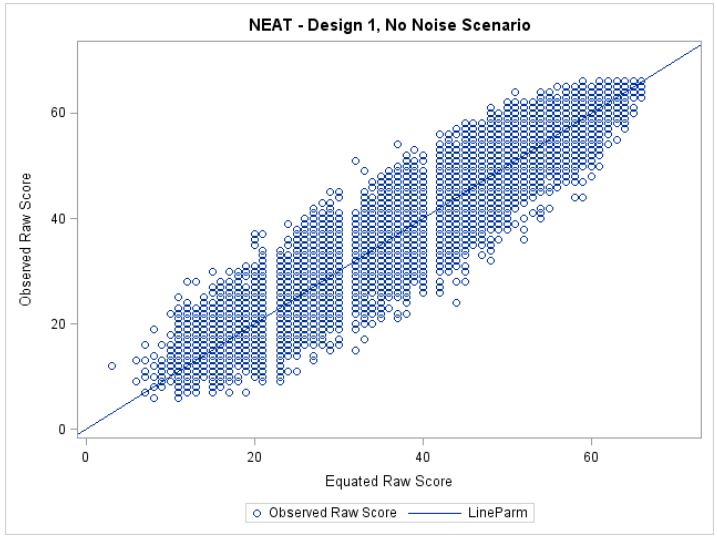

Figure 9 Plot of x and $\hat{y}$, NEAT Design 1 Ideal

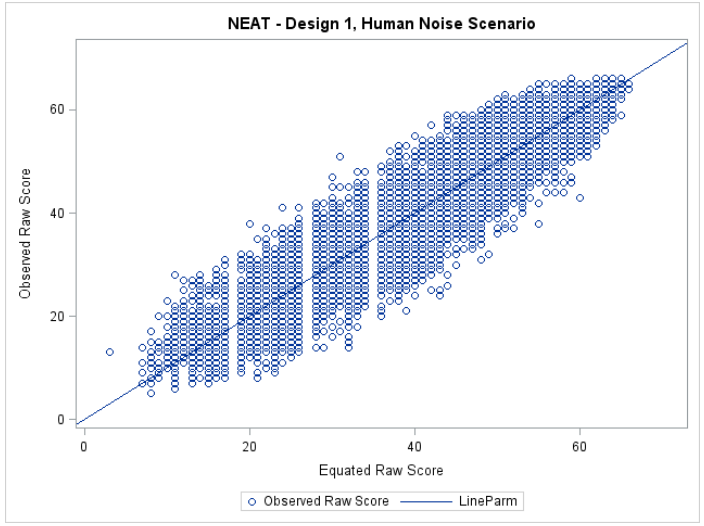

Figure 10 Plot of x and $\hat{y}$, NEAT Design 1 Human

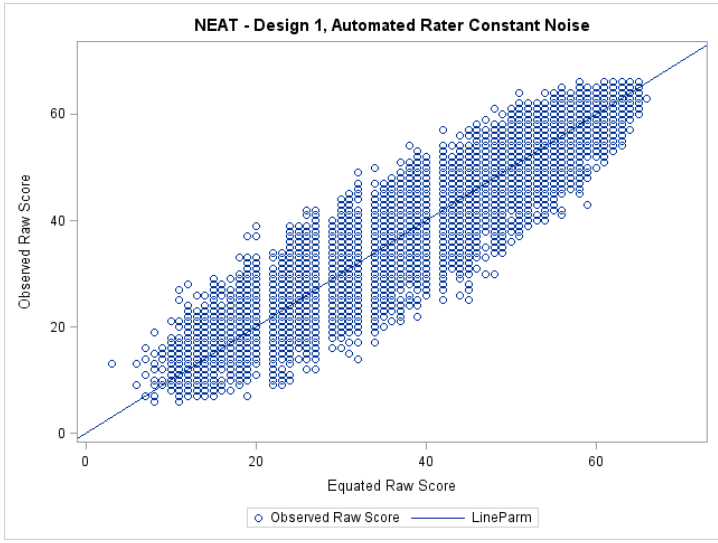

Figure 11 Plot of x and $\hat{y}$, NEAT Design 1 Automated Rater (Constant Noise)

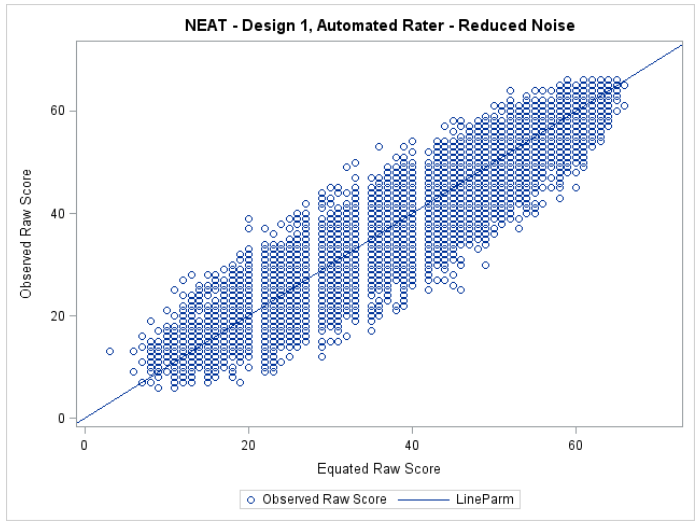

Figure 12 Plot of x and $\hat{y}$, NEAT Design 1 Automated Rater (Reduced Noise)

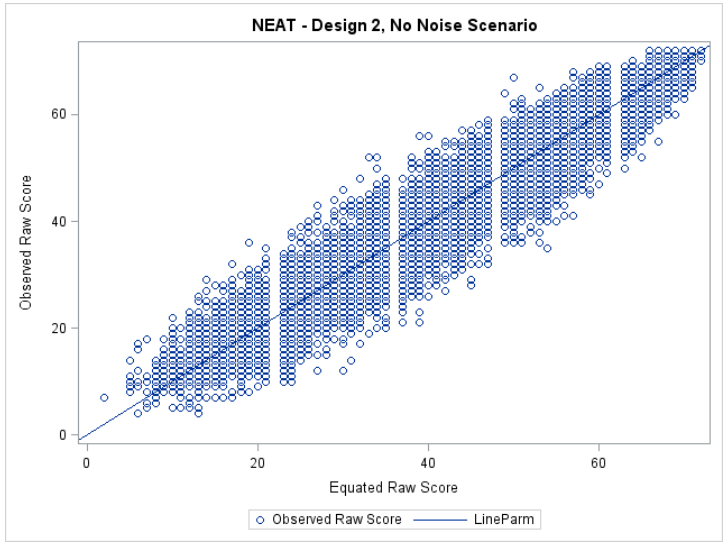

Figure 13 Plot of x and $\hat{y}$, NEAT Design 2 Ideal

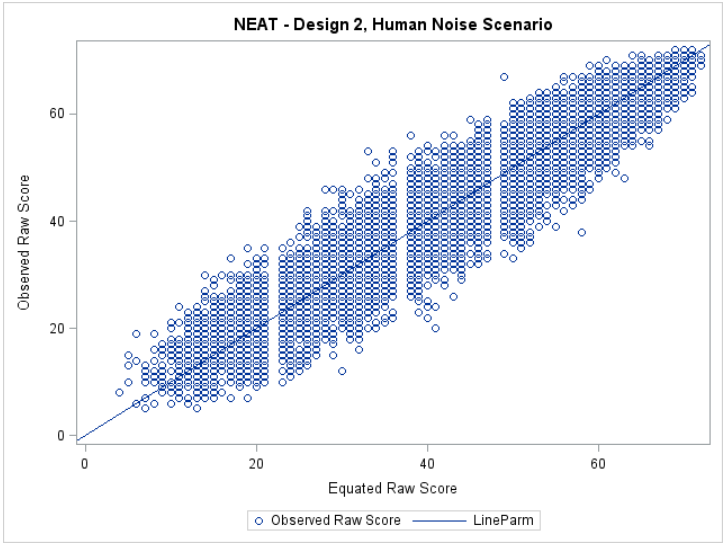

Figure 14 Plot of x and $\hat{y}$, NEAT Design 2 Human

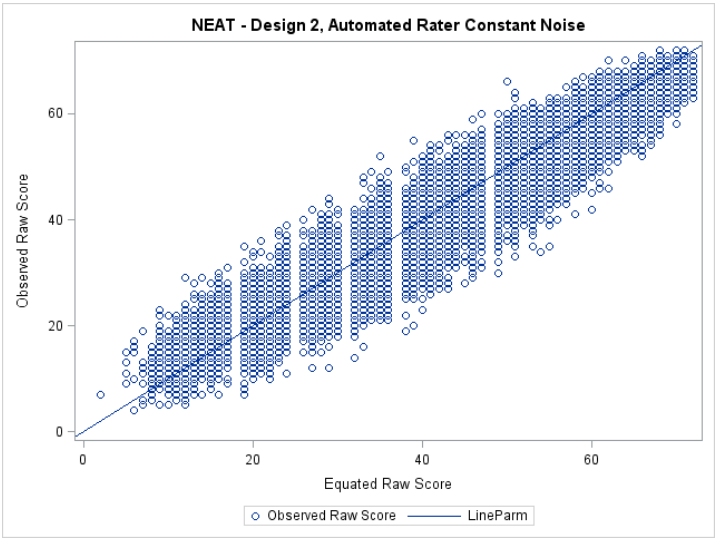

Figure 15 Plot of x and $\hat{y}$, NEAT Design 2 Automated Rater (Constant Noise)

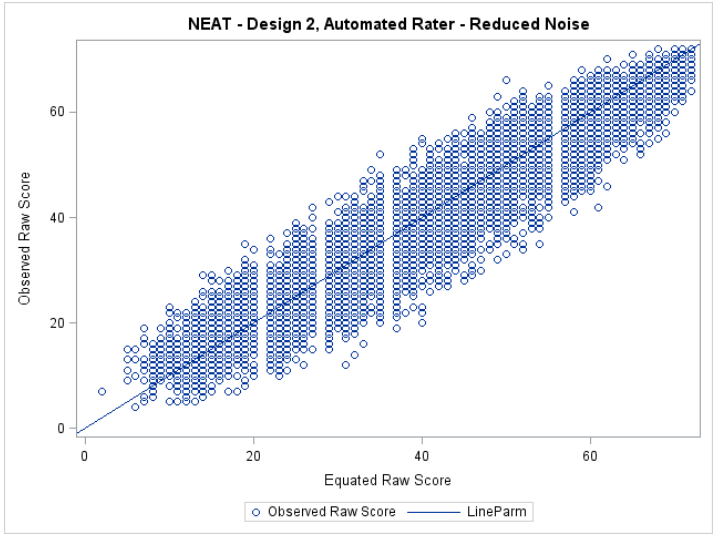

Figure 16 Plot of x and $\hat{y}$, NEAT Design 2 Automated Rater (Reduced Noise)

4.3 Examinee Impact

The impact of rater bias and variability on examinee ability estimates ($\hat{\theta}$) was also examined. The $\hat{\theta}$ values based on equated test forms were compared between each noise scenario and the ideal rater (scenarios as described in Section 3.3.3). In this way, the differences in $\hat{\theta}$ that are attributable entirely to the amount of rater bias and variability added can be evaluated directly. The results show that RMSE values are notably smaller than for the raw score comparisons noted in Section 4.1, ranging from 0.03 to 0.20 for single group and 0.05 to 0.10 for NEAT. Nevertheless, the same pattern that very slightly favors the automated rater scenario, where rater variability was reduced in form 2 under both test designs, is noted. In all cases, the RMSE values for the human and automated rater (constant noise) scenarios are about double that resulting from analysis of the automated rater scenario where rater bias and variability were reduced.

Looking at the correlations between $\hat{\theta}$ values based on noisy and ideal raters, the patterns show a very strong relationship for all scenarios, although there is a small, but noticeable difference for design 2 under the human rater scenario, single group equating design. Where all other correlations round to 1.00, the human to ideal scenario under single group design 2 is 0.97, and the automated rater with constant bias and variability to ideal raters is 0.98. Refer to Tables 8 and 9 for detailed RMSE and correlation results.

Table 8 RMSE and Correlation between $\hat{\theta}$ for Human and Ideal, Single Group

	RMSE		r	
Rater Scenario	Design 1	Design 2	Design 1	Design 2
Human & Ideal	0.06	0.20	0.997	0.974
Automated (Constant Noise) & Ideal	0.06	0.18	0.998	0.980
Automated (Reduced Noise) & Ideal	0.03	0.10	0.999	0.993

*Design 1 contains 1 of 3 constructed responses. Design 2 contains 4 of 6 constructed response items for the Constant Noise Scenario

Table 9 RMSE and Correlation between $\hat{\theta}$ for Human and Ideal, NEAT

	RMSE		r	
Rater Scenario	Design 1	Design 2	Design 1	Design 2
Human & Ideal	0.08	0.11	0.996	0.994
Automated (Constant Noise) & Ideal	0.05	0.11	0.999	0.994
Automated (Reduced Noise) & Ideal	0.05	0.06	0.999	0.998

Figures 17-28 show these relationships graphically, providing a view of the differences along the full raw score scales for each design. The $\hat{\theta}$ correspondence is shown to be as close at the correlations suggest they would be. For example, based on the RMSE values in Table 8, the expectation is that the single group, human and automated rater (constant noise) scenarios under test design 2 would produce scatter plots with a slightly wider spread of correspondence. Referring to Figure 17-22, this pattern is noted. However, it is further noted that the pattern is not consistent along the raw score scale. In fact, there is a very tight correspondence in the center of the scale, but there is more variation at the lower and upper ends of the scale. These patterns are consistent with the raw score comparisons in Section 4.2. Additionally, there is what appears to be a slight tendency to under-predict equated scores at the upper, and to over-predict at the lower end. Close inspection of the plots for all remaining scenarios, across both equating designs, shows that this pattern is often repeated although to a lesser degree in the NEAT results.

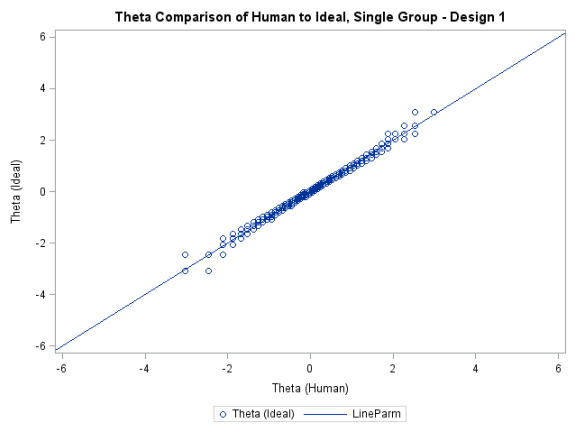

Figure 17 $\widehat{\theta}$ Comparison Human v Ideal Single Group, D1

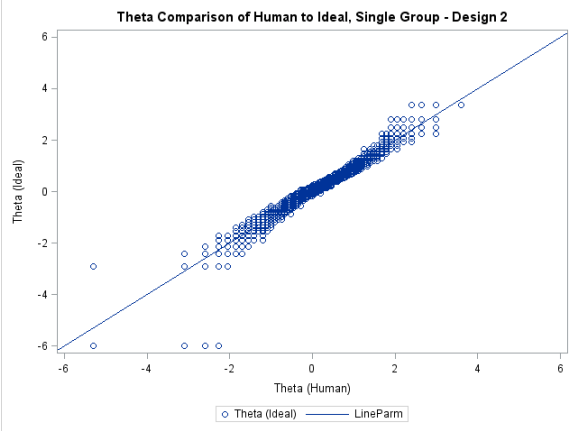

Figure 20 $\widehat{\theta}$ Comparison Human v Ideal, Single Group, D2

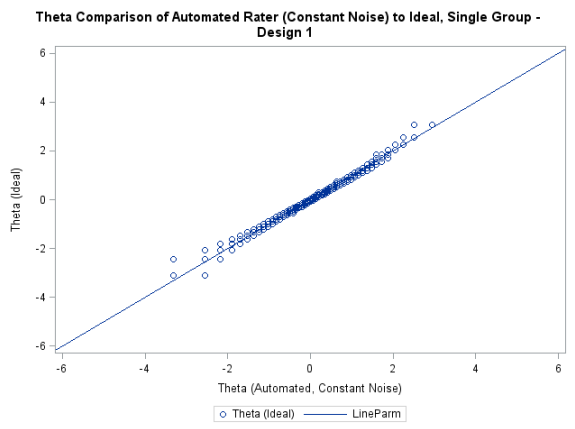

Figure 18 $\widehat{\theta}$ Comparison, Automated Rater (Constant Noise) v Ideal, Single Group, D1

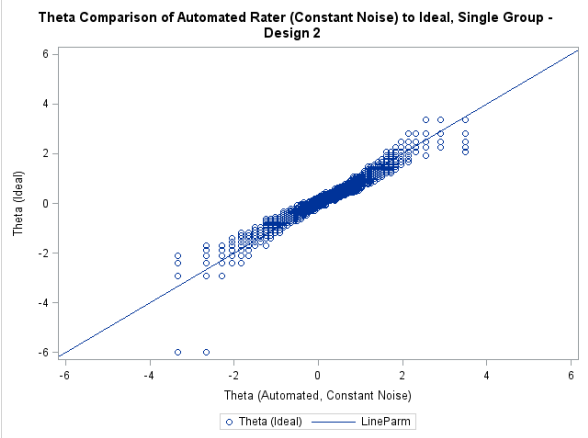

Figure 21 $\widehat{\theta}$ Comparison, Automated Rater (Constant Noise) v Ideal, Single Group, D2

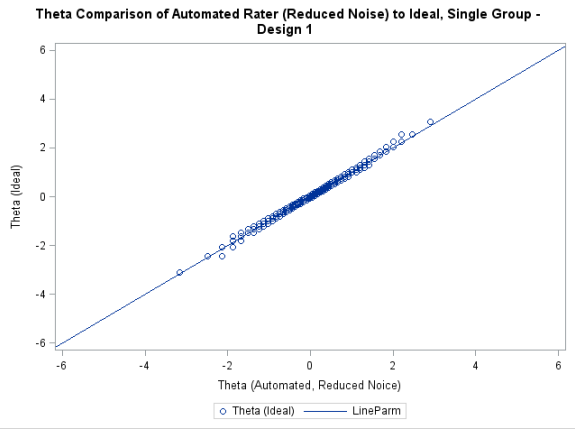

Figure 19 $\widehat{\theta}$ Comparison, Automated Rater (Reduced Noise) v Ideal, Single Group, D1

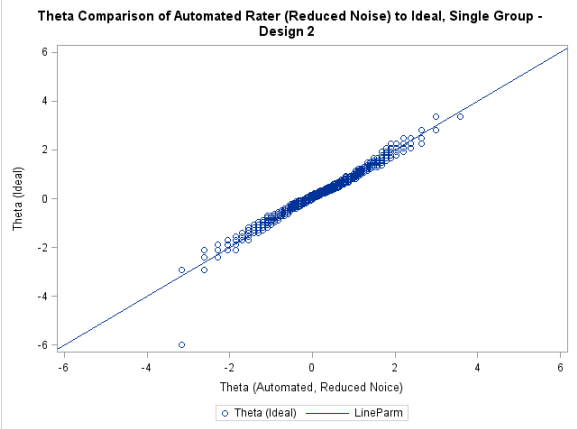

Figure 22 $\widehat{\theta}$ Comparison, Automated Rater (Reduced Noise) v Ideal, Single Group, D2

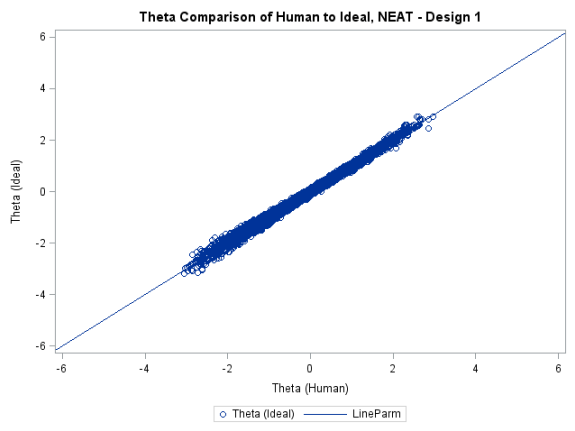

Figure 23 $\hat{\theta}$ Comparison Human v Ideal NEAT, D1

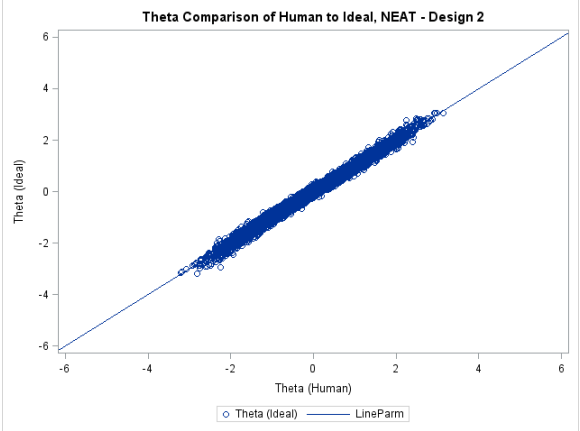

Figure 26 $\hat{\theta}$ Comparison Human v Ideal, NEAT, D2

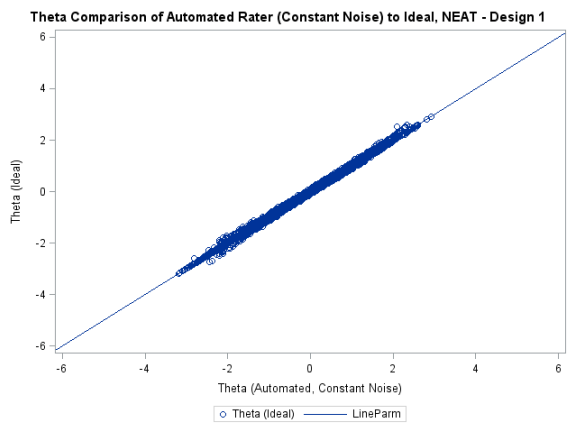

Figure 24 $\hat{\theta}$ Comparison, Automated Rater (Constant Noise) v Ideal, NEAT, D1

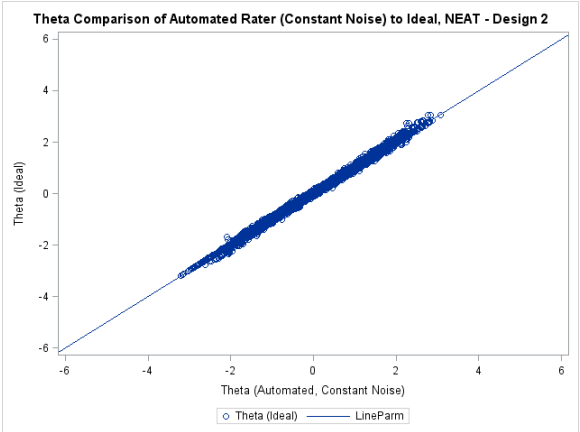

Figure 27 $\hat{\theta}$ Comparison, Automated Rater (Constant Noise) v Ideal, NEAT, D2

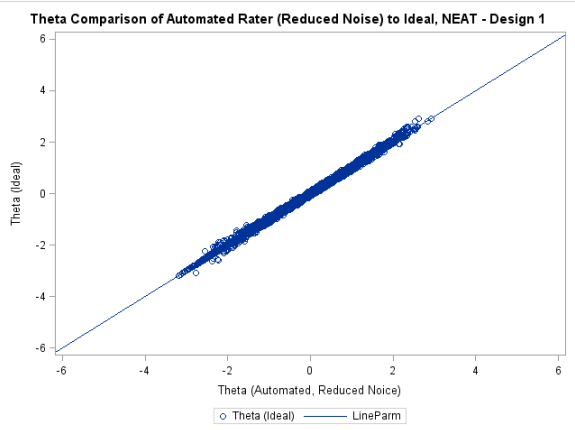

Figure 25 $\hat{\theta}$ Comparison, Automated Rater (Reduced Noise) v Ideal, NEAT, D1

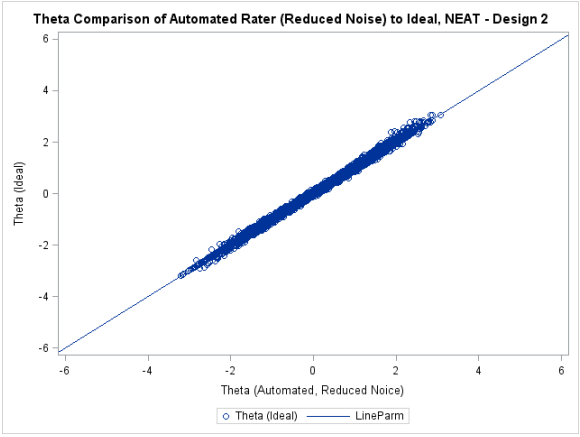

Figure 28 $\hat{\theta}$ Comparison, Automated Rater (Reduced Noise) v Ideal, NEAT, D2

To understand the potential impact that these rater noise scenarios have on classification decisions where the most examinees are located, a single cut was set in the center of the distributions. Performance level classification rates between ideal scores and those based on noisy raters were then compared (See Table 10). Results show very small impact in the center of the examinee ability distributions, at < 5% maximum. For the single group equating results, and in particular, for test design 2, there is a more systematic impact at the defined decision point as the bias resulting from the equating appears to occur across a larger portion of this scale than for the other scenarios. In these cases, more examinees who pass under the ideal scenario would fail when the rater noise is present. For example, it is noted that 4.20% of examinees who pass under ideal rater conditions, would fail under the human noise condition. Only 0.19% of examinees who failed under the ideal scenarios would pass under the human rater scenario. This result indicates that the downward equating bias noted in Figures 20-22 extends from the upper to the center of the scale in this scenario, more so than in the other scenarios examined here.

However, based on the general locations of the bias effects, namely in the tails of the score scales, larger bias effects would be expected for cuts set at the lower (upward bias) and upper (downward bias) ends of the score scale. Based on the patterns noted Figures 17-28, the rater noise simulated here would have a larger impact on performance level decisions made at these locations. Also, comparative inferences about examinees, e.g. percentile ranks or growth measures, based on scores in these regions of the scales may be threatened. Refer to Table 10 for detailed performance level comparisons.

Table 10 Performance Level Impact where $\widehat{\boldsymbol{\theta}} = \mathbf{0}$

Design	Rater	Percent Perfect Agreement	Percent Discrepant where Ideal = Fail	Percent Discrepant where Ideal = Pass
Single Group, Design 1	Human (Constant Noise)	99.18	0.38	0.44
	Automated (Constant Noise)	99.28	0.29	0.43
	Automated (Reduced Noise)	99.86	0.03	0.11
Single Group, Design 2	Human (Constant Noise)	95.61	0.19	4.20
	Automated (Constant Noise)	96.29	0.62	3.09
	Automated (Reduced Noise)	96.98	0.08	2.94
NEAT, Design 1	Human (Constant Noise)	98.11	1.00	0.89
	Automated (Constant Noise)	98.77	0.62	0.61
	Automated (Reduced Noise)	98.98	0.53	0.52
NEAT, Design 2	Human (Constant Noise)	97.21	1.34	1.45
	Automated (Constant Noise)	98.88	0.60	0.52
	Automated (Reduced Noise)	98.82	0.66	0.52

4.3 IRT Equating

Stocking and Lord (1983) results for the NEAT equating design are summarized next. Table 11 shows the correlation of anchor item parameters (multiple-choice and combined) for each test design. Two correlations stand out in test design 2, under the automated rater scenarios. The item discrimination parameter correlation between forms 1 and form 2, is noticeably lower than the other correlations in these cases. Table 12 provides the mean a- and b-parameter values for the linking items, which are strongly influenced by the inclusion of the multiple-choice anchor items as well. Overall mean discrimination is consequently quite similar across all scenarios, although the human rater scenario shows a slight degradation of discrimination for test design 2. Similarly, the influence of the multiple-choice anchor items on mean difficulty results in very small differences across scenarios, where changes from the ideal appear to be larger for the automated rater scenarios, particularly for test design 2.

Table 11 Anchor Item Correlations, S&L

	Ideal Design 1	Ideal Design 2	Human Design 1	Human Design 2	Automated Design 1 (Constant Noise)	Automated Design 2 (Constant Noise)	Automated Design 1 (Reduced Noise)	Automated Design 2 (Reduced Noise)
a-par	0.932	0.976	0.949	0.982	0.896	0.734	0.903	0.730
b-par	0.998	0.999	0.994	0.999	0.996	0.991	0.996	0.992
c-par	0.947	0.993	0.953	0.990	0.947	0.994	0.952	0.994

Table 12 Mean Discrimination and Difficulty of Anchor Item Parameters

	IRT Item Parameter Linking Item Means by Design and Form							
	a-Parameter				b-Parameter			
Rater Scenario	D1 F1	D1 F2	D2 F1	D2 F2	D1 F1	D1 F2	D2 F1	D2F2
Ideal	1.039	1.096	1.052	1.056	-0.249	-0.113	0.055	0.028
Human	0.933	1.053	0.925	0.924	-0.196	-0.113	0.059	0.035
Automated (Constant Noise)	0.993	1.080	0.939	1.021	-0.196	-0.094	0.147	0.055
Automated (Reduced Noise)	0.933	1.077	0.939	1.017	-0.196	-0.100	0.147	0.060

*D=Design; F=Form

The Stocking and Lord constants and minimized loss functions are documented in Table 13. A notable result is that the minimum loss function values are generally larger for design 2 with 6 versus 3 constructed response items, as might be expected.

Table 13 Stocking & Lord Equating Constants and Minimum Loss Function

	Ideal Design 1	Ideal Design 2	Human Design 1	Human Design 2	Automated Design 1 (Constant Noise)	Automated Design 2 (Constant Noise)	Automated Design 1 (Reduced Noise)	Automated Design 2 (Reduced Noise)
A	1.03	1.02	1.04	1.02	1.10	1.17	1.10	1.11
B	-0.13	0.02	-0.08	0.01	-0.08	0.06	-0.08	0.06
Minimized Loss function	0.0033	0.0093	0.0035	0.0103	0.0094	0.0045	0.0087	0.0134

Test characteristic curves for forms 1 and 2, linking item sets in Forms 1 and 2, and the equated test for each rater bias and variability scenario and test design are provided in Figures 29-36. These plots provide a graphical summary of NEAT equating results under each bias scenario. The results show that the anchor item set selected is slightly more difficult than either form 1 (test1) or 2 (test2) in design 1. Overall, small differences are

noted between the curves for the original and rescaled forms for all rater noise scenarios for design 1.

Upon closer inspection of where differences occur across ideal, human, and automated rater scenarios however, an interesting pattern is noted in terms of where along the TCCs the differences are greatest. Comparing the ideal to the human rater scenario, there appears to be some impact of the rater bias and variability effect on the *a*- and *b*-parameters. Namely, the human rater scenario resulted in better alignment of all five curves in terms of difficulty, but both the human and automated rater scenarios resulted in less discriminating equated scores. The results for the automated rater (reduced noise) scenario shows a similar alignment of the 5 curves versus the ideal with slightly more degradation in the discrimination of the equated scores.

Looking at the results for test design 2, the same patterns are noted, although the degradation in discrimination of the equated test scores is steeper, and this is particularly true for the automated rater scenario where the rater bias and variability are held constant across the 2 forms. It appears that the influence of the levels of rater bias and unreliability on item parameters can shift test difficulty around in unpredictable ways, but that the discrimination is systematically degraded by the presence of consistently greater levels of bias and variability than by reducing bias and variability.

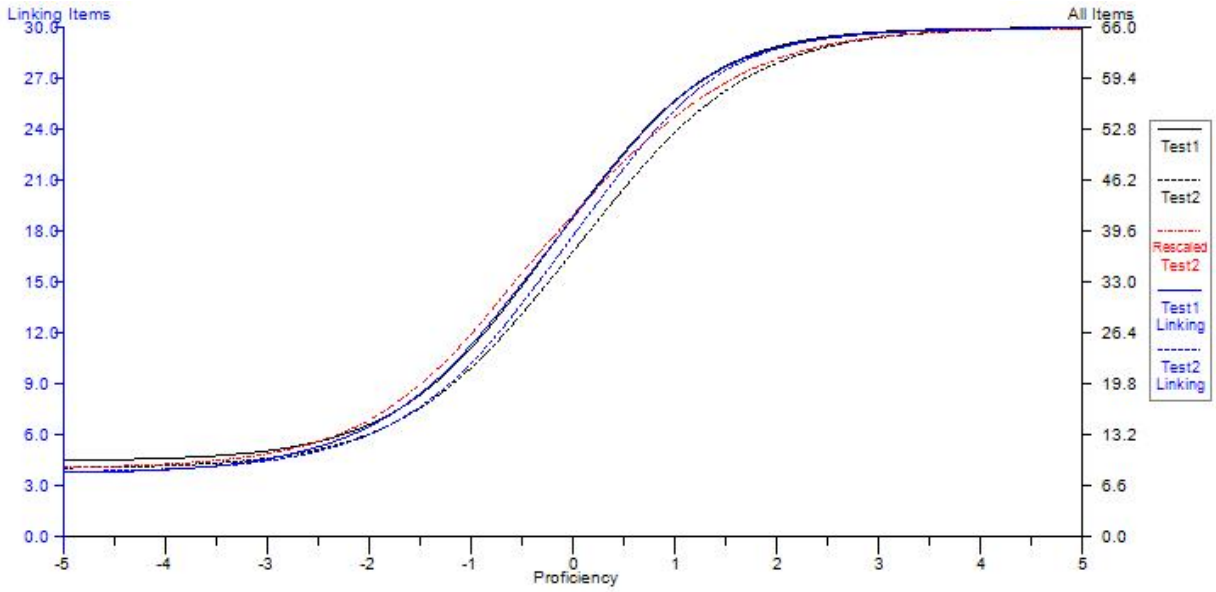

Figure 29 TCC Comparison, Ideal Rater Scenario Design 1

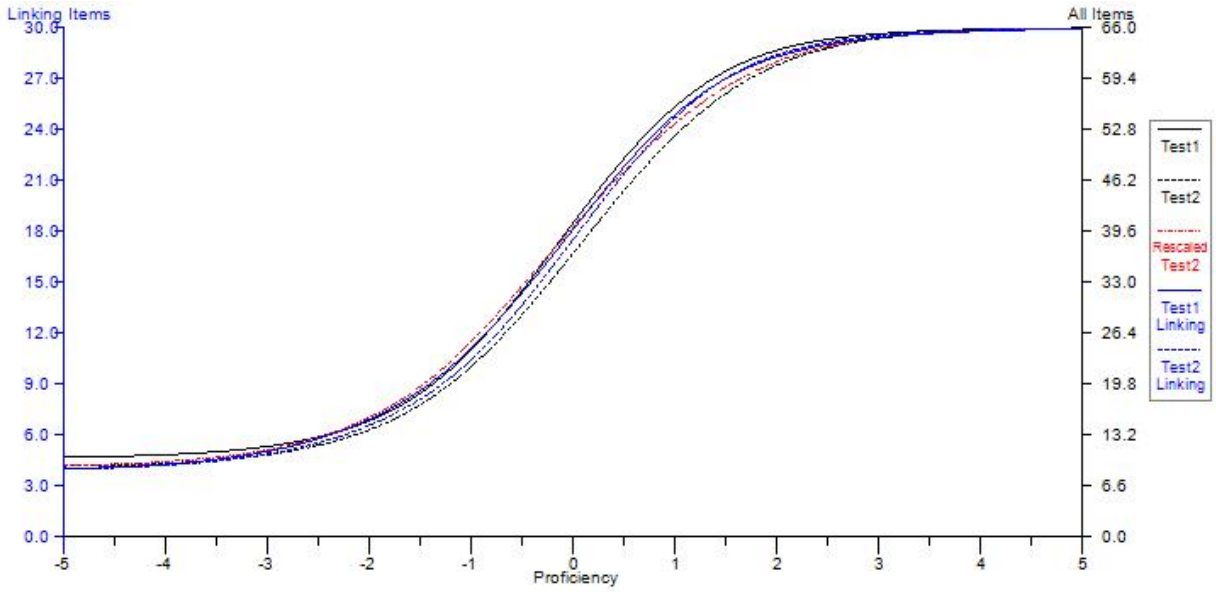

Figure 30 TCC Comparison, Human Rater Scenario Design 1

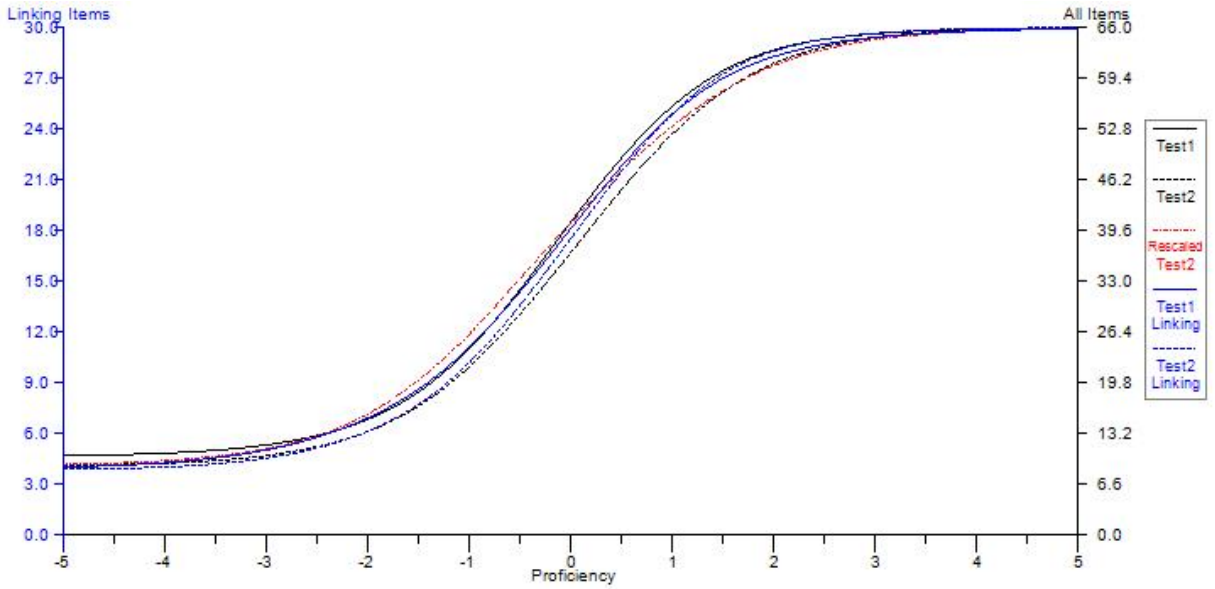

Figure 31 TCC Comparison, Automated Rater (Constant Noise) Scenario Design 1

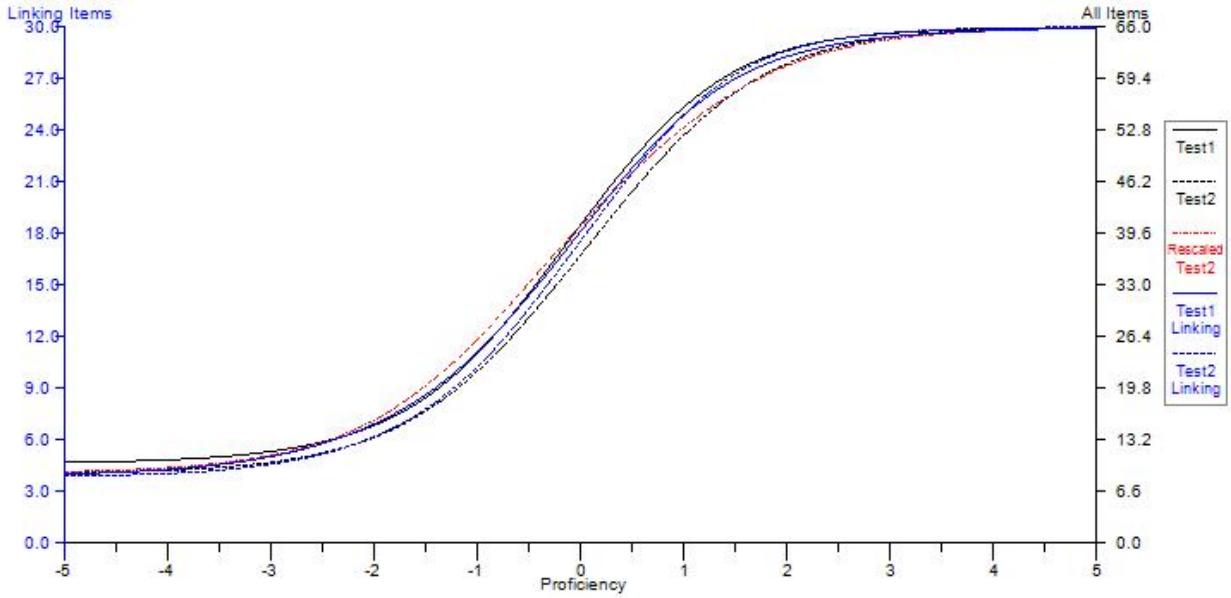

Figure 32 TCC Comparison, Automated Rater (Reduced Noise) Scenario Design 1

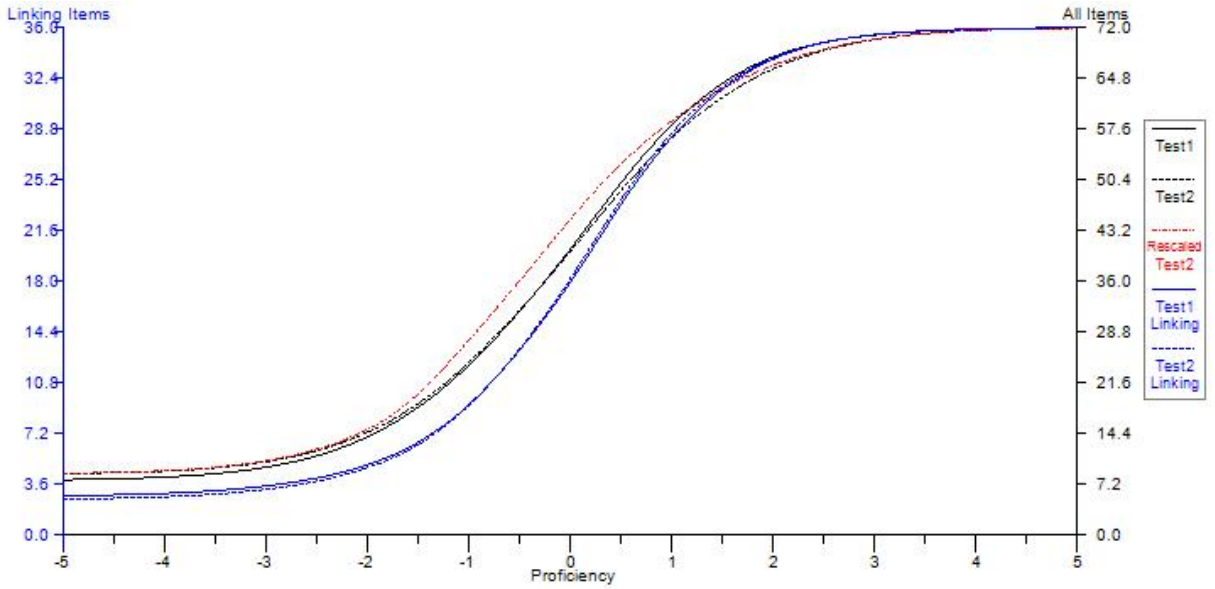

Figure 33 TCC Comparison, Ideal Scenario Design 2

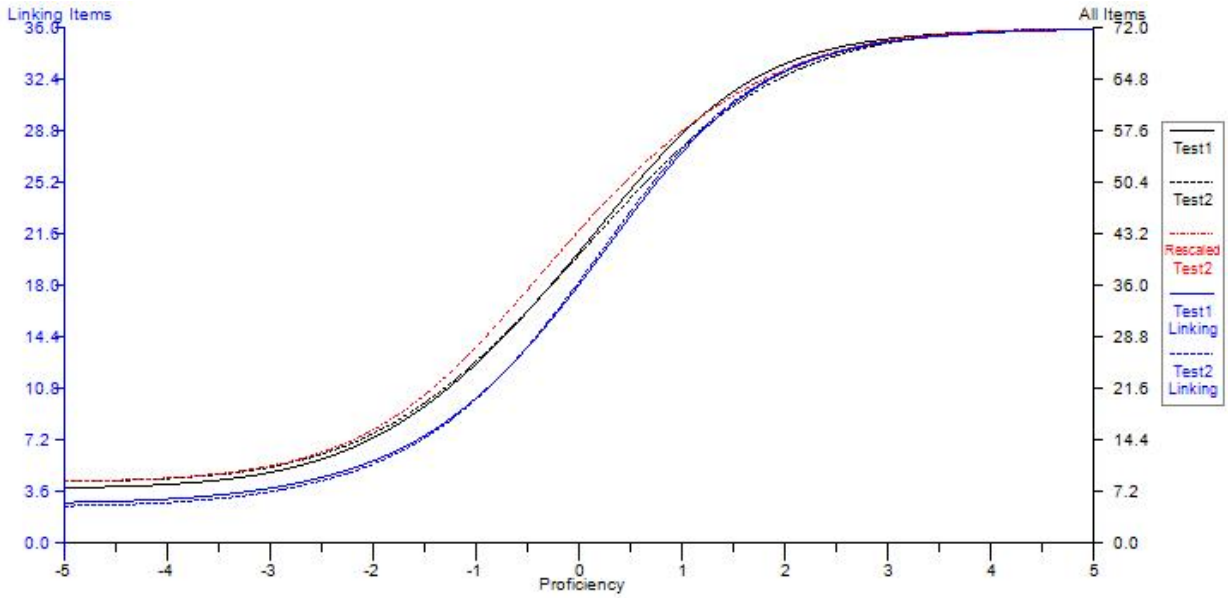

Figure 34 TCC Comparison, Human Scenario Design 2

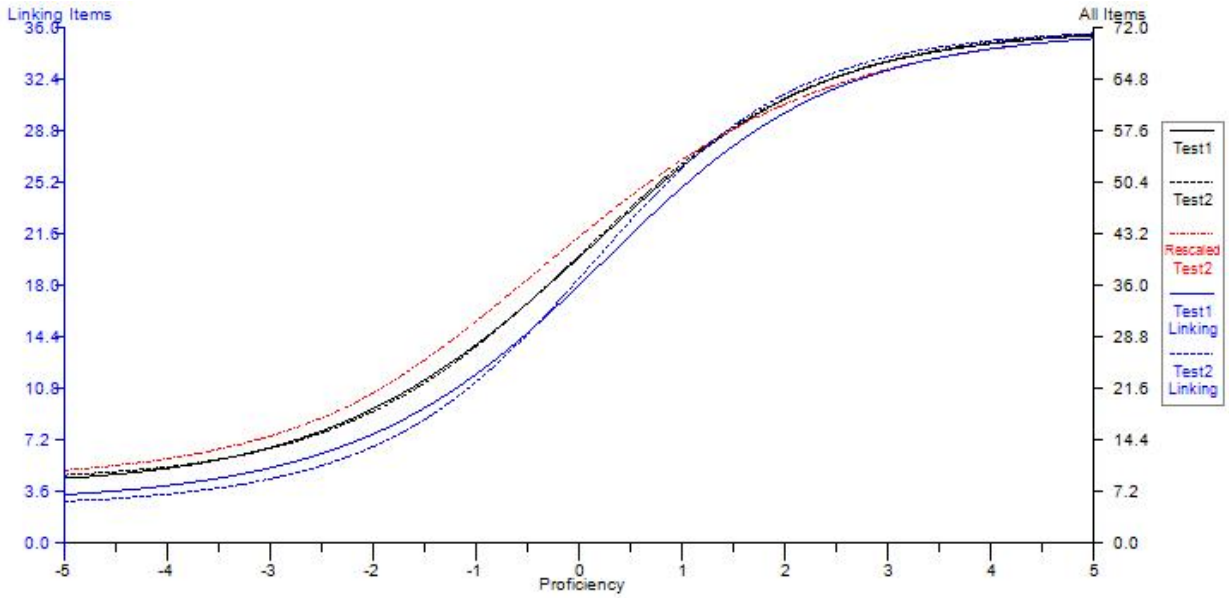

Figure 35 TCC Comparison, Automated Rater (Constant Noise) Scenario Design 2

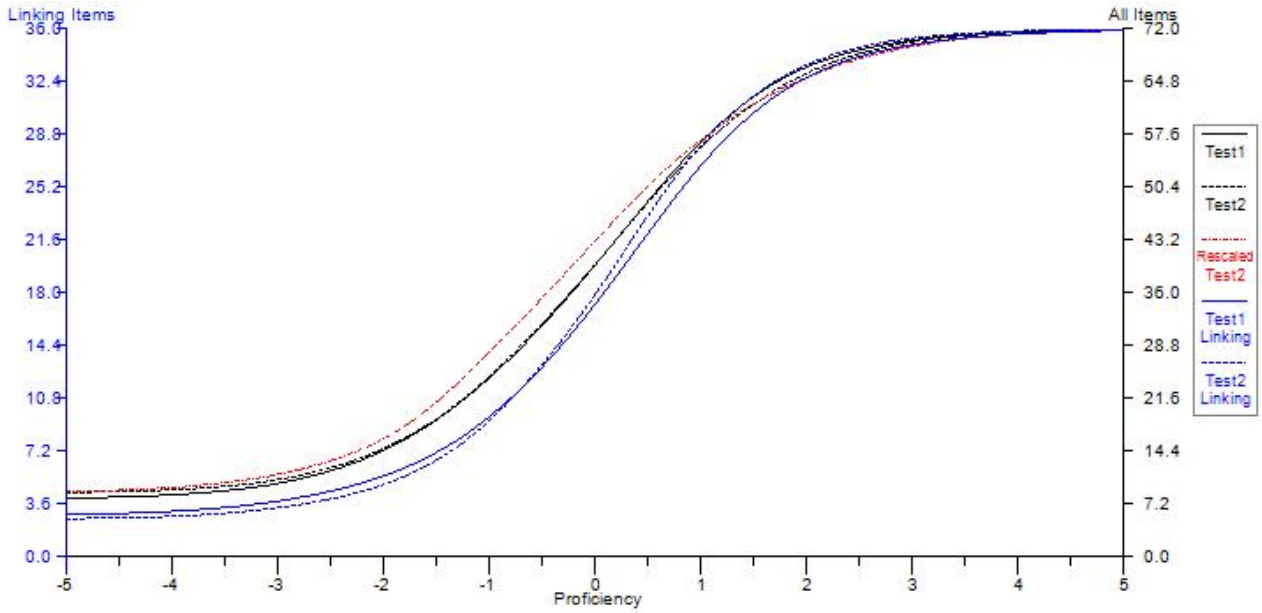

Figure 36 TCC Comparison, Automated Rater (Reduced Noise) Scenario Design

CHAPTER V

DISCUSSION

5.1 Review of Study Purpose, Method of Investigation, and Research Questions

This study examines the potential impact of different levels of systematic rater error on test score equity, and on the inferences and decisions made based on those scores. The primary research question that this study addresses is, what are the effects of rater bias and variability on test equating solutions? More specifically, and for the purpose of directly addressing this main question, the following questions are posed: 1) What are the effects of changes in automated rater bias on test equating solutions? 2) Do changes in rater bias change the inferences that can be made about test scores that are intended to be comparable? 3) What is the impact to examinee scores and performance level classifications under these conditions? 4) And, are either common item or common examinee designs more robust to changes in rater bias?

To provide context for evaluating different rater bias scenarios, with automated raters as the primary focus of this study, typical equating solutions based on human raters with and without bias were investigated to provide a baseline understanding of the effect of human rater noise and noise-free ratings on test equating solutions, under two test designs, and under two equating procedures. Further, a scenario was included whereby a theoretically lowest level of acceptable automated rater quality was held constant across test forms to provide a baseline for understanding the impact of improving rater quality on score equity across forms.

The simulated "test forms" were scaled and equated under these four rater conditions, ideal, human, and two automated rater scenarios where bias and variability were held

constant across the forms, and where rater bias and variability were reduced. The results were evaluated in two main ways. First, the equated raw scores (i.e. predicted form 1 scores) were compared to the observed form 1 scores. In this way, the impact of rater noise on equated raw scores could be examined relative to a theoretically ideal rater scenario, while attempting to isolate the effect of the selected levels of rater bias and variability.

Second, $\hat{\theta}$ comparisons were examined between human and ideal rater scenarios, and between automated and ideal rater scenarios. The evaluation of $\hat{\boldsymbol{\theta}}$ allowed an investigation of the examinee impact, which, as a consequence of the rater noise on equated results is important to understand. As the ultimate motivation for testing is to use examinee scores for some purpose, it is important to understand the potential comparability of decisions made about individuals under different rater error scenarios. To the extent that examinees are classified differently based on different equating outcomes, such comparability could be threatened. Likewise, where decisions are made based on differences between examinees along the scale, it is important to understand the size of the impact of rater bias and variability on estimates of examinee ability at any point on the scale.

5.2 Effects of Changes in Automated Rater Bias on Test Equating Solutions and Score Comparability

A primary research question for this study is, what are the effects of changes in automated rater bias on test equating solutions? To distinguish automated raters from human raters, the HRM model was used to introduce values of rater bias and variability that are consistent with findings in a study reporting such values both human and automated raters (Casabianca et al., 2016). Additionally, once the lowest acceptable level of automated rater noise was established, two scenarios for automated raters were derived—one where the high

end of rater bias and variability (i.e. the lowest acceptable state of art) was held constant, and another where bias and variability are reduced by 50%. It is, therefore, important to note that the manner in which human and automated raters are distinguished in this study is just one way that such raters might be characterized, and that many other rater bias and variability values could be identified.

For example, in this study, the automated rater had slightly higher bias values than the average human rater. This may not necessarily be true in all cases. Similarly, automated and human raters are not likely to rate all 2-, 3-, and 4-point items with the same amount of bias and variability, as was implemented here where the levels were different across rater scenarios, but the same within rater scenarios, within item max score type (refer to Table 3). Raters are often trained individually on items and any number of interactions between the rater, the item, and the examinee response may trigger more or less bias and variability in an item. Nevertheless, the rater scenarios selected for this study represent a reasonable starting point, grounded in empirical results of rater quality analyses, for an investigation of this research question.

The root mean square error and correlation values between equated form 2 and observed form 1 scores were evaluated and results indicate that the human and the automated rater (constant noise) rater scenarios produced consistently slightly higher RMSEs than the automated rater (reduced noise) scenario, suggesting that lower rater bias and variability, even where it changes across forms has a smaller effect than where bias and variability are constant. The differences in RMSE values across scenarios were quite small, but this pattern was consistent. The correlations between the equated and observed scores were very strong and generally indistinguishable across scenarios.

To better understand impact across the raw score scales, equated form 2 raw scores and form 1 observed scores were plotted in Figures 1-16. These figures show a slight tendency for rater noise to bias scores downwards at the upper end of the scale and upwards at the lower end. Although these effects appear to be small, they are systematic across test and equating designs, and across rater scenarios. Importantly, the effect is more notable for test design 2 where there are 6 versus 3 constructed response items are in the test design, noting that the single group design is not directly comparable across equating scenarios as only one constructed response item was used for design 1 under the single group equating, and that the automated rater (constant bias) scenario contained 4 of the 6 constructed response items for design 2 under the single group equating.

So, in terms of the impact of changes in automated rater bias and variability on test equating solutions, when compared with ideal, human, and consistently applied automated rater bias and variability, differences indicate that the automated rater scenario with reducing bias and variability simulated here produces results closer to an ideal criterion than the human or automated (constant noise) rater scenario. This suggests a possibility that decisions to use improved automated raters in equating programs may be better supported than anticipated under test designs and conditions similar to those simulated in this study.

The bias noted in the equating results appears quite small, but in the strictest sense, this does indicate a potential threat to the comparability of the inferences made about scores at the upper and lower ends of the raw score scales, particularly under the single group design. However, the differences noted under these test designs and conditions may be small enough to be practically insignificant, particularly in cases where classification decisions are made in more typical locations along the scale, e.g. toward the center and not in the tails.

More caution may be warranted for use of percentiles and growth measures that intend to draw meaning from score differences at more extreme locations along the scale.

5.4 Impact to examinee scores and performance level classifications

An examination of the differences in examinee ability shows a similar impact to the raw score analyses. Where rater variability is higher under the simulated human and automated (constant noise) rater scenarios, there is a higher level of discordance between examinee $\hat{\theta}$s, particularly for the single group results. Referring to Figures 17-28, the correspondence between results based on noisy and ideal ratings is consistently better for automated raters where reduced noise in form 2 is applied, and generally, for the test design with fewer constructed response items. However, similar to the raw score discrepancy patterns, the correspondence between $\hat{\theta}$s based on ideal and noisy ratings shows a noticeable systematic impact at the higher and lower end of the scale, particularly for test design 2 for human and automated (constant noise) rater scenarios, under the single group equating design.

Looking at the impact of setting a cut score in the center of the scale, maximum misclassification of examinees is less than 5% for all rater error scenarios, so the impact to examinees under any of these scenarios may actually be quite small, where cut scores are set in typical locations away from extreme locations along the scale. As mentioned in the preceding discussion, however, there would be more cause for concern for decisions made based on score differences in the extreme ends of the scale, e.g. through the use of a cut score set in these regions, and through the use of percentiles or growth measures. Ultimately the decision to use an improved automated rater under conditions similar to those simulated here, would need to consider the decisions being made about examinees, where along the scale the

decision point exists, and how much bias effect in the equating is considered too much for the purposes of the assessement. In this study, the maximum misclassification of examinees due to bias in the equating was most often under 2% which may be small enough to provide acceptable results in some situations. Caution would be warranted in situations where more extreme consequences are attached to score results, e.g. high school graduation, college entrance, or professional certifications, particularly for decisions made at the upper and lower ends of the scores ranges.

5.5 Use of Common Examinee and Common Item Equating Designs

Finally, based on the RMSE and correlation results between equated form 2 and form 1 observed raw scores, and between $\hat{\theta}$s comparing each noise scenario to the ideal, it does not appear that either the single group or NEAT equating are definitively more robust to changes in rater bias and variability across equated forms. However, the patterns of correspondence do suggest that the noted bias effects are at least slightly larger for the single group equating design, particularly for design 2. It is important to note that 2 items in the single group design failed in their IRT parameter estimation for test design 1, and 2 failed for the automated rater (constant bias) in test design 2. This may be significant in that the expectation, based on the overall results, would be a reduction in the number of items used in the equating would result in less equating bias, all other conditions held equal. That fewer items in the single group equating, resulted in greater equating bias for some scenarios, suggests that NEAT equating design may indeed be more robust to the presence of, and change in rater bias.

5.6 Conclusions, Limitations, and Future Studies

Rater bias and variability have what appears to be a predictable impact on item discrimination, but not on item difficulty. In both single group and NEAT IRT equating scenarios, results show some bias in the equating results, notably in the tails of the raw score scales used here, and more notably in the single group, test design 2 rater noise scenarios. The implication for examinee impact is that more caution in interpretation of results may be warranted where decisions are made at the higher or lower ends of the scale, including the use of scores to compute growth or make other examinee comparisons at the lower or upper ends of the scale.

Reducing bias and variability in the equated forms appears to produce a slightly less biased equating result then holding the values constant at a higher rate across both forms. This suggests that, possibly, rater improvement scenarios may be preferable to holding the rater quality constant in a test equating program. This finding is unexpected due to long held assumptions that equity would be violated if raters perform differently on the same items across testing occasions. Previous studies, for example Tate (1999 & 2000) and Kim, Walker, and McHale (2010a & 2010b), conclude that bias in test equating solutions is reduced when constructed response anchor item difficulties are adjusted based on rescoring procedures. In such cases, examinee responses on a base form are scored by rater groups A and B, where group B is responsible for scoring examinee responses on the equate form. The differences between the two sets of scores are then used to adjust item difficulty on the equated form prior to equating.

In cases where multiple-choice and constructed response item scores are not highly correlated, the use of such adjustments, sometimes referred to as “rater drift” adjustments, are

shown to result in less equating bias than using either a multiple-choice only anchor, or not performing a drift adjustment at all. However, rater performance was not directly measured in these studies, and only the differences in group A and B ratings were used to establish the drift adjustments. The Kim et al studies assumed, as this study assumed at the start, that a change in rater bias violates score equity, but they did not account for the level of systematic rater error present. It is possible the rater bias and variability present were much smaller than those use in the current study. Kim et al. (2010a & 2010b) also provide a view of the impact to equating error, showing that the unadjusted scenario actually resulted in lower equating error, even where their selected measure of equating bias showed a reduction.

The bias noted in the equated results for the current study does appear to be influenced by larger numbers of constructed response items used in the equating. This is consistent with other findings that conclude the use of constructed response items in equating may result in greater equating error, particularly where the correlations between multiple-choice and constructed response items are lower than desired (Dorans, 2004; Hagge and Kolen, 2012; Kim and Walker, 2009). Consequently, results for different test and equating designs might draw very different conclusions about the level of rater noise that is acceptable in a selected rater, and how much of a reduction across test forms might be tolerated under equity considerations. In the scenarios examined here, some decisions, particularly in the center of the score scale, would see very small performance level differences between forms 1 and 2. This appears to be more true for test designs with fewer constructed response items.

Ultimately, these results appear to indicate that systematic rater error, as defined in this study, may represent a relatively small threat to the comparability of the inferences that can be made about the equated scores, and that more caution is warranted in equity

assumptions where decisions are made based on scores at the extremes of the score scales. It should be noted that, although the human rater bias and variability were held constant across items and across forms in this study, this may not necessarily be true empirically. This choice was made primarily for the purpose of contrasting an approximation of the reality that we generally accept some bias in human ratings, and that using an average of bias and variability from the Casabianca et al. (2016) study could reasonably represent such an approximation. That is a strong assumption, but one made frequently when calibrating tests while treating the item level scores as ideal.

In reality, some raters may be harsh on one item and lenient on another, so the approach taken here may present a conservative (i.e., pessimistic) view of the cumulative impact of bias and variability across items. Examination of more complex patterns of rater bias should be investigated. Nevertheless, as the noise levels used for a baseline automated rater quality in this study were intended to approximate a minimally acceptable operational model, these findings suggest that holding systematic error at a constant at a theoretically highest acceptable level is possibly a greater threat to score equity then improving rater precision across equated forms.

A general limitation of this study, then, is that the selected noise scenarios may not generalize to all that might be observed under different test and equating designs, IRT distribution parameters, and parameter estimation, and equating methods. The levels of bias chosen here represent a small sample of the possible scenarios that might be simulated, even when constraining the levels to correspond to realistic automated rater agreement scenarios, e.g. where the noise results in 0.70 or greater QWK in agreement between ideal and noisy ratings. Replication over more varied test design, examinee ability distributions, and bias and

variability scenarios will be important to test the generalizability of these results. Also, additional evaluation of the quality of these equating results will be important. For example, the invariance of examinee classifications based theta scores over equated forms will be important to understand the impact for testing scenarios where examinee classifications are used.

Finally, there is a clear need to test the methods employed here on empirical data sets to more fully examine the question of impact of changing rater bias and variability across equated forms in operational settings. Although these findings provide some initial insight into the impact of typical automated rater bias on test equating, they are ultimately limited by the level of correspondence between the simulation parameters used here and those that exist empirically across different testing designs and populations.

5.6 Summary of Study Importance

This work was motivated by the psychometric context of continued improvements in the state of the art for automated scoring. Such improvements are driven by demands to demonstrate the efficacy of automated scoring models and procedures, in terms of both score precision and validity. Just as the quality of human rater scores is a concern for the psychometric properties of tests, so is the quality of automated rater scores, and because improvement is the goal, the impact of such changes in rater quality needs to be better understood.

This study focused specifically on the impact that improvements in automated raters might have on test equating solutions, where no studies have been identified that have examined this question. The scenarios examined here start with an automated rater that minimally meets current industry standards for rater agreement with a theoretically ideal

rater, i.e. a QWK of no lower than 0.70. It then examines the impact on equating solutions under the conditions of holding rater quality constant, and of improvement where improvement is defined as reducing rater bias and variability by half. The equating results for each of these scenarios are compared to each other, to a human rater scenario, and to a theoretical criterion of based on ideal ratings.

Findings suggest that both the single group and NEAT equating designs tolerate the levels of rater bias and variability simulated here reasonably well, although equated scores at the extreme ends of the score scale show some bias, and the NEAT equating design appears to result in slightly lower levels of bias in the equated scores. This is an important finding as it is reasonable to assume scoring improvements in large scale testing programs are desirable, and that if such improvements do not impose unacceptable levels of bias in equated solutions, then rater improvements may not unduly threaten our assumptions about score equity under conditions that are similar to those simulated here. Further, if replications across additional scenarios and empirical results conclude similarly, we are further motivated to reduce rater bias and variability in automated (and human) raters, as we may not be presented with the need to make a choice between greater precision and equity. This is potentially good news for assessment stakeholders that would like to take advantage of the perceived benefits of implementing automated scoring solutions, namely the potentially reduced costs and increased scoring speeds that are desired.

APPENDIX

ITEM PARAMETERS

Table A1 IRT Item Parameters, Single Group, Design 1, Ideal Raters, Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0499	-1.3861	0.1009
MC2	2	3PLM	1.0815	1.0022	0.1167
MC3	2	3PLM	1.0614	-1.1607	0.0858
MC4	2	3PLM	1.0838	-1.3128	0.2490
MC5	2	3PLM	1.0253	0.3339	0.0623
MC6	2	3PLM	0.9094	0.8825	0.1586
MC7	2	3PLM	1.3555	-1.0130	0.1692
MC8	2	3PLM	0.7365	0.9170	0.1889
MC9	2	3PLM	0.8895	0.7473	0.1238
MC10	2	3PLM	1.1040	-0.1572	0.0284
MC11	2	3PLM	0.9306	-0.7301	0.2747
MC12	2	3PLM	0.8592	-0.7075	0.1109
MC13	2	3PLM	0.7803	1.3515	0.2478
MC14	2	3PLM	1.0590	-0.8647	0.2524
MC15	2	3PLM	1.0441	-1.9225	0.1620
MC16	2	3PLM	0.5486	-0.5397	0.1983
MC17	2	3PLM	0.8167	1.1614	0.1767
MC18	2	3PLM	0.9582	0.2869	0.0486
MC19	2	3PLM	0.8357	2.4358	0.2865
MC20	2	3PLM	1.0649	0.3360	0.0593
MC21	2	3PLM	1.0793	-0.9518	0.2872
MC22	2	3PLM	1.0047	1.1466	0.0091
MC23	2	3PLM	1.1518	-0.8209	0.1134
MC24	2	3PLM	0.9478	-0.5356	0.0585
MC25	2	3PLM	1.2488	0.4727	0.1307
MC26	2	3PLM	0.9033	-0.4186	0.2899
MC27	2	3PLM	0.8083	0.2093	0.2109
MC28	2	3PLM	1.1512	-0.0239	0.2237
MC29	2	3PLM	1.5057	0.3156	0.0542
MC30	2	3PLM	1.2072	-0.0024	0.1258
MC31	2	3PLM	1.3337	-0.0338	0.0179
MC32	2	3PLM	1.2179	-1.4863	0.1601
MC33	2	3PLM	0.9261	0.7822	0.1531
MC34	2	3PLM	1.3205	-0.0666	0.0808
MC35	2	3PLM	1.1768	0.3554	0.1642
MC36	2	3PLM	0.9683	-0.7957	0.0978

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC37	2	3PLM	1.0071	-1.4573	0.2768
MC38	2	3PLM	0.8227	1.6885	0.0355
MC39	2	3PLM	0.9352	-0.5018	0.0654
MC40	2	3PLM	1.1793	0.4714	0.2788
MC41	2	3PLM	0.9932	-0.6167	0.2156
MC42	2	3PLM	0.9900	-0.4755	0.2249
MC43	2	3PLM	0.8343	1.5913	0.2558
MC44	2	3PLM	0.8137	0.3322	0.2812
MC45	2	3PLM	1.0974	0.8891	0.0919
MC46	2	3PLM	1.0457	0.1626	0.2342
MC47	2	3PLM	0.9038	1.4117	0.1748
MC48	2	3PLM	0.8924	0.3861	0.0521
MC49	2	3PLM	1.0337	-0.7465	0.0805
MC50	2	3PLM	0.9453	0.0445	0.2636
MC51	2	3PLM	1.4143	1.5052	0.2413
MC52	2	3PLM	1.6877	-1.2257	0.1285
MC53	2	3PLM	0.9345	0.0536	0.1699
MC54	2	3PLM	1.0688	1.4953	0.3154
MC55	2	3PLM	0.8993	0.3090	0.2736
MC56	2	3PLM	1.0341	-1.2373	0.1463
MC57	2	3PLM	0.9702	0.0630	0.1748
FR1	3	GPCM	1.1122	0.3879	

Table A2 Equated IRT Item Parameters, Single Group, Design 1, Ideal Raters, Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0811	-1.4772	0.1750
MC2	2	3PLM	0.7366	2.7203	0.0881
MC3	2	3PLM	0.9535	0.3522	0.1473
MC4	2	3PLM	0.9541	-0.1498	0.2170
MC5	2	3PLM	1.1482	0.0976	0.0957
MC6	2	3PLM	0.9824	1.5738	0.3182
MC7	2	3PLM	1.3105	1.7759	0.0226
MC8	2	3PLM	1.1122	-0.3956	0.1933
MC9	2	3PLM	0.8963	-0.8962	0.1449
MC10	2	3PLM	1.4458	0.0990	0.2457
MC11	2	3PLM	1.0351	-0.1234	0.2417
MC12	2	3PLM	0.9433	2.1900	0.2928
MC13	2	3PLM	1.2356	-0.5742	0.1051
MC14	2	3PLM	0.9924	-0.3539	0.1923
MC15	2	3PLM	1.0155	-0.6690	0.2976
MC16	2	3PLM	1.4228	0.0346	0.2646
MC17	2	3PLM	1.7456	1.7668	0.2407
MC18	2	3PLM	1.3027	-0.1286	0.0323
MC19	2	3PLM	1.2184	-1.0805	0.2577
MC20	2	3PLM	1.1400	-0.4235	0.1994
MC21	2	3PLM	1.1045	-0.0061	0.1740
MC22	2	3PLM	1.1680	0.7933	0.1732
MC23	2	3PLM	1.2485	-0.5615	0.2333
MC24	2	3PLM	0.9931	0.4755	0.2580
MC25	2	3PLM	1.1149	0.6358	0.0415
MC26	2	3PLM	0.9680	-1.2771	0.2292
MC27	2	3PLM	0.9543	0.8575	0.2344
MC28	2	3PLM	1.0838	1.0062	0.1148
MC29	2	3PLM	1.2852	1.1292	0.0806
MC30	2	3PLM	0.9820	0.7928	0.1293
MC31	2	3PLM	1.2317	-0.4644	0.2479
MC32	2	3PLM	0.9219	0.2634	0.0670
MC33	2	3PLM	1.2686	1.4115	0.0501
MC34	2	3PLM	0.9736	-0.0100	0.1859
MC35	2	3PLM	0.9110	-0.2199	0.2432
MC36	2	3PLM	1.5739	0.0443	0.1752
MC37	2	3PLM	1.2597	-0.5904	0.0626
MC38	2	3PLM	0.9485	-0.1267	0.1540
MC39	2	3PLM	1.2119	-1.0554	0.3019
MC40	2	3PLM	1.0962	-0.3942	0.3153

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	0.9464	-1.0818	0.2284
MC42	2	3PLM	0.8737	-1.4136	0.1191
MC43	2	3PLM	0.9852	0.0675	0.1160
MC44	2	3PLM	1.2427	0.2581	0.0850
MC45	2	3PLM	0.7970	-2.2829	0.2314
MC46	2	3PLM	1.3000	-0.0173	0.2923
MC47	2	3PLM	0.8274	0.2442	0.1931
MC48	2	3PLM	1.2249	0.3633	0.0283
MC49	2	3PLM	0.8139	-1.0055	0.1095
MC50	2	3PLM	0.9937	1.6797	0.2491
MC51	2	3PLM	0.8456	0.0736	0.2321
MC52	2	3PLM	1.1031	0.3623	0.2830
MC53	2	3PLM	0.8746	0.9159	0.0892
MC54	2	3PLM	0.5737	-0.2925	0.1801
MC55	2	3PLM	0.8888	0.7700	0.2263
MC56	2	3PLM	0.7415	-1.3112	0.2172
MC57	2	3PLM	1.2138	1.1164	0.1540
FR1	3	GPCM	0.8820	0.5521	

Table A3 IRT Item Parameters, Single Group, Design 2, Ideal Raters, Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.6784	-1.0546	0.1775
MC2	2	3PLM	0.8484	-0.1508	0.1166
MC3	2	3PLM	0.6600	0.5361	0.0824
MC4	2	3PLM	0.7503	1.4874	0.1037
MC5	2	3PLM	1.0131	0.5395	0.2321
MC6	2	3PLM	1.0627	-0.1920	0.2164
MC7	2	3PLM	1.6391	2.3338	0.2255
MC8	2	3PLM	1.2121	-0.4815	0.0505
MC9	2	3PLM	1.0276	-2.2885	0.2315
MC10	2	3PLM	0.9952	-0.5707	0.0989
MC11	2	3PLM	1.3271	1.6627	0.1193
MC12	2	3PLM	1.0675	-1.6342	0.1960
MC13	2	3PLM	1.4329	0.2751	0.1010
MC14	2	3PLM	1.3270	0.5714	0.2543
MC15	2	3PLM	0.8874	-0.7527	0.1437
MC16	2	3PLM	0.8306	-0.5381	0.1956
MC17	2	3PLM	1.0134	0.5720	0.2333
MC18	2	3PLM	1.2301	0.5466	0.2940
MC19	2	3PLM	0.8171	0.6166	0.1576
MC20	2	3PLM	1.0948	-1.3376	0.2254
MC21	2	3PLM	1.1472	0.5749	0.1196
MC22	2	3PLM	0.5719	-0.2250	0.1115
MC23	2	3PLM	1.3821	0.8620	0.2551
MC24	2	3PLM	0.9760	0.3704	0.2038
MC25	2	3PLM	0.7634	0.5741	0.2040
MC26	2	3PLM	1.1980	1.2934	0.2186
MC27	2	3PLM	1.2300	-0.1686	0.1891
MC28	2	3PLM	1.4469	0.5730	0.0518
MC29	2	3PLM	0.9459	-0.6201	0.2487
MC30	2	3PLM	0.8723	2.3002	0.2080
MC31	2	3PLM	1.1415	-1.6250	0.1806
MC32	2	3PLM	0.9314	0.7415	0.0764
MC33	2	3PLM	1.0600	-0.1114	0.2164
MC34	2	3PLM	1.1523	1.1644	0.1663
MC35	2	3PLM	0.7144	2.0595	0.2084
MC36	2	3PLM	1.2423	-1.4202	0.0926
MC37	2	3PLM	0.9038	-0.8269	0.2386
MC38	2	3PLM	1.1970	-0.8767	0.2736
MC39	2	3PLM	1.0086	0.7393	0.2470
MC40	2	3PLM	0.7756	0.5342	0.2496

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	0.9676	-0.4502	0.0650
MC42	2	3PLM	1.4973	0.5405	0.0666
MC43	2	3PLM	0.8170	-2.8890	0.1962
MC44	2	3PLM	0.9957	1.0712	0.1961
MC45	2	3PLM	1.2124	-1.2146	0.1641
MC46	2	3PLM	1.0433	-1.2891	0.1931
MC47	2	3PLM	0.7455	0.1033	0.1278
MC48	2	3PLM	1.0333	0.9173	0.0520
MC49	2	3PLM	0.9161	-0.2859	0.2383
MC50	2	3PLM	0.9660	-0.1666	0.2487
MC51	2	3PLM	0.8197	1.0476	0.0594
MC52	2	3PLM	1.0436	-0.2609	0.1897
MC53	2	3PLM	1.4060	-0.2671	0.2294
MC54	2	3PLM	1.1479	-0.9839	0.1908
FR1	3	GPCM	0.9509	0.2702	
FR2	3	GPCM	0.7425	-0.1856	
FR3	4	GPCM	0.9685	-0.1208	
FR4	4	GPCM	1.0136	-0.1586	
FR5	5	GPCM	1.1922	0.4217	
FR6	5	GPCM	0.9002	-0.5133	

Table A4 Equated IRT Item Parameters, Single Group, Design 2, Ideal Raters, Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.9523	1.4782	0.0229
MC2	2	3PLM	1.2820	-1.4648	0.1546
MC3	2	3PLM	0.9711	1.6601	0.0380
MC4	2	3PLM	0.8115	-0.2619	0.2557
MC5	2	3PLM	0.9089	0.4542	0.0373
MC6	2	3PLM	1.1355	-0.2464	0.0508
MC7	2	3PLM	1.1525	0.7692	0.0756
MC8	2	3PLM	1.1925	0.5801	0.2349
MC9	2	3PLM	0.8315	0.6576	0.0442
MC10	2	3PLM	1.2172	0.1729	0.0512
MC11	2	3PLM	0.9064	1.1048	0.0308
MC12	2	3PLM	1.1626	0.8911	0.0490
MC13	2	3PLM	0.8622	-0.0305	0.1378
MC14	2	3PLM	0.8653	1.1984	0.2145
MC15	2	3PLM	0.8520	-1.3295	0.3179
MC16	2	3PLM	1.0495	-0.5639	0.1857
MC17	2	3PLM	1.2790	0.8400	0.2742
MC18	2	3PLM	1.1668	-0.6594	0.0430
MC19	2	3PLM	1.1308	-1.8292	0.2018
MC20	2	3PLM	1.0585	0.8137	0.1070
MC21	2	3PLM	1.2339	-0.6758	0.0853
MC22	2	3PLM	0.8901	1.2588	0.1577
MC23	2	3PLM	1.3468	1.5449	0.1599
MC24	2	3PLM	0.8135	0.0191	0.0384
MC25	2	3PLM	1.0695	0.2098	0.0384
MC26	2	3PLM	0.8583	0.3215	0.1132
MC27	2	3PLM	1.1156	1.0216	0.0298
MC28	2	3PLM	0.8369	0.0045	0.2542
MC29	2	3PLM	1.1174	0.6716	0.1476
MC30	2	3PLM	1.3502	1.2323	0.2651
MC31	2	3PLM	0.6577	-0.7946	0.0849
MC32	2	3PLM	0.8642	-0.7934	0.0912
MC33	2	3PLM	1.0265	-0.5105	0.2437
MC34	2	3PLM	0.9521	-1.4389	0.2161
MC35	2	3PLM	1.0150	0.3473	0.0415
MC36	2	3PLM	0.9501	-0.6867	0.1014
MC37	2	3PLM	0.8421	-0.4305	0.3220
MC38	2	3PLM	0.7235	-0.2280	0.2792
MC39	2	3PLM	1.1124	-1.0510	0.1250
MC40	2	3PLM	0.9714	0.7753	0.1849

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	0.8353	2.7844	0.0280
MC42	2	3PLM	0.9869	1.9085	0.1667
MC43	2	3PLM	0.6100	0.0898	0.2354
MC44	2	3PLM	1.0253	-0.3482	0.1928
MC45	2	3PLM	0.8918	-0.5214	0.2363
MC46	2	3PLM	1.0358	-1.9860	0.1450
MC47	2	3PLM	0.9023	0.0147	0.2691
MC48	2	3PLM	0.8345	0.0081	0.2279
MC49	2	3PLM	1.3850	1.0988	0.1527
MC50	2	3PLM	1.0509	0.3436	0.1447
MC51	2	3PLM	0.6265	0.4663	0.0952
MC52	2	3PLM	0.8083	-0.0586	0.0981
MC53	2	3PLM	0.6925	-1.0617	0.2471
MC54	2	3PLM	0.8611	0.6642	0.2403
FR1	3	GPCM	0.7504	0.8742	
FR2	3	GPCM	0.5959	0.6985	
FR3	4	GPCM	1.1837	0.4470	
FR4	4	GPCM	1.1836	-0.2613	
FR5	5	GPCM	0.7956	0.6622	
FR6	5	GPCM	1.4238	-0.8708	

Table A5 IRT Item Parameters, Single Group, Design 1, Human Raters, Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0402	-1.4261	0.0883
MC2	2	3PLM	1.0477	0.9969	0.1138
MC3	2	3PLM	1.0519	-1.2024	0.0795
MC4	2	3PLM	1.0504	-1.4028	0.2164
MC5	2	3PLM	1.0170	0.3246	0.0654
MC6	2	3PLM	0.8871	0.8674	0.1544
MC7	2	3PLM	1.3354	-1.0474	0.1691
MC8	2	3PLM	0.7280	0.9285	0.1939
MC9	2	3PLM	0.8805	0.7370	0.1247
MC10	2	3PLM	1.0802	-0.1808	0.0290
MC11	2	3PLM	0.9229	-0.7295	0.2890
MC12	2	3PLM	0.8545	-0.7506	0.1028
MC13	2	3PLM	0.7849	1.3585	0.2532
MC14	2	3PLM	1.0480	-0.8936	0.2536
MC15	2	3PLM	1.0411	-1.9600	0.1491
MC16	2	3PLM	0.5296	-0.6153	0.1807
MC17	2	3PLM	0.7952	1.1652	0.1766
MC18	2	3PLM	0.9565	0.2797	0.0523
MC19	2	3PLM	0.8304	2.4233	0.2849
MC20	2	3PLM	1.0367	0.3109	0.0567
MC21	2	3PLM	1.0390	-1.0179	0.2729
MC22	2	3PLM	0.9822	1.1431	0.0085
MC23	2	3PLM	1.1289	-0.8667	0.0993
MC24	2	3PLM	0.9398	-0.5590	0.0621
MC25	2	3PLM	1.2525	0.4572	0.1312
MC26	2	3PLM	0.8867	-0.4457	0.2906
MC27	2	3PLM	0.7791	0.1678	0.2014
MC28	2	3PLM	1.1207	-0.0586	0.2171
MC29	2	3PLM	1.4865	0.2988	0.0546
MC30	2	3PLM	1.1829	-0.0226	0.1247
MC31	2	3PLM	1.3168	-0.0474	0.0189
MC32	2	3PLM	1.2124	-1.5344	0.1471
MC33	2	3PLM	0.9089	0.7675	0.1504
MC34	2	3PLM	1.3081	-0.0798	0.0842
MC35	2	3PLM	1.1545	0.3386	0.1615
MC36	2	3PLM	0.9577	-0.8391	0.0904
MC37	2	3PLM	0.9797	-1.5538	0.2391
MC38	2	3PLM	0.8065	1.6942	0.0341
MC39	2	3PLM	0.9222	-0.5349	0.0609
MC40	2	3PLM	1.1736	0.4600	0.2799

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	0.9656	-0.6818	0.1987
MC42	2	3PLM	0.9525	-0.5450	0.2048
MC43	2	3PLM	0.8492	1.6047	0.2618
MC44	2	3PLM	0.7898	0.3141	0.2765
MC45	2	3PLM	1.0596	0.8781	0.0896
MC46	2	3PLM	1.0438	0.1451	0.2328
MC47	2	3PLM	0.8836	1.4134	0.1754
MC48	2	3PLM	0.8813	0.3813	0.0553
MC49	2	3PLM	1.0149	-0.7885	0.0711
MC50	2	3PLM	0.9313	0.0100	0.2577
MC51	2	3PLM	1.3997	1.5037	0.2424
MC52	2	3PLM	1.6600	-1.2703	0.1255
MC53	2	3PLM	0.9123	0.0229	0.1655
MC54	2	3PLM	1.0428	1.4894	0.3151
MC55	2	3PLM	0.8901	0.2979	0.2730
MC56	2	3PLM	1.0240	-1.2727	0.1489
MC57	2	3PLM	0.9588	0.0481	0.1774
FR1	3	GPCM	0.4420	0.5447	

Table A6 Equated IRT Item Parameters, Single Group, Design 1, Human Raters, Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0710	-1.5296	0.1591
MC2	2	3PLM	1.2907	2.1478	0.0677
MC3	2	3PLM	0.9271	0.3338	0.1421
MC4	2	3PLM	0.9414	-0.1848	0.2115
MC5	2	3PLM	1.1194	0.0716	0.0905
MC6	2	3PLM	1.0007	1.5790	0.3202
MC7	2	3PLM	1.2816	1.7873	0.0218
MC8	2	3PLM	1.0934	-0.4162	0.1984
MC9	2	3PLM	0.8829	-0.9350	0.1360
MC10	2	3PLM	1.4274	0.0742	0.2453
MC11	2	3PLM	1.0122	-0.1617	0.2339
MC12	2	3PLM	0.9391	2.1960	0.2946
MC13	2	3PLM	1.2499	-0.5866	0.1158
MC14	2	3PLM	0.9638	-0.3917	0.1892
MC15	2	3PLM	1.0034	-0.7041	0.2975
MC16	2	3PLM	1.3965	0.0118	0.2641
MC17	2	3PLM	1.8475	1.7832	0.2420
MC18	2	3PLM	1.2966	-0.1440	0.0378
MC19	2	3PLM	1.1894	-1.1238	0.2557
MC20	2	3PLM	1.1208	-0.4484	0.1982
MC21	2	3PLM	1.0832	-0.0314	0.1721
MC22	2	3PLM	1.1468	0.7907	0.1758
MC23	2	3PLM	1.2177	-0.6079	0.2227
MC24	2	3PLM	0.9893	0.4661	0.2603
MC25	2	3PLM	1.0933	0.6236	0.0426
MC26	2	3PLM	0.9785	-1.2771	0.2486
MC27	2	3PLM	0.9581	0.8537	0.2350
MC28	2	3PLM	1.0740	1.0074	0.1155
MC29	2	3PLM	1.2683	1.1253	0.0808
MC30	2	3PLM	0.9764	0.7946	0.1330
MC31	2	3PLM	1.1971	-0.5121	0.2357
MC32	2	3PLM	0.9015	0.2460	0.0646
MC33	2	3PLM	1.2410	1.4165	0.0482
MC34	2	3PLM	0.9491	-0.0487	0.1763
MC35	2	3PLM	0.9054	-0.2187	0.2544
MC36	2	3PLM	1.5182	0.0163	0.1693
MC37	2	3PLM	1.2249	-0.6387	0.0526
MC38	2	3PLM	0.9120	-0.1746	0.1427
MC39	2	3PLM	1.1655	-1.1351	0.2766
MC40	2	3PLM	1.0636	-0.4348	0.3052

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	0.9196	-1.1612	0.2010
MC42	2	3PLM	0.8646	-1.4335	0.1324
MC43	2	3PLM	0.9678	0.0424	0.1136
MC44	2	3PLM	1.2388	0.2516	0.0891
MC45	2	3PLM	0.8073	-2.2970	0.2209
MC46	2	3PLM	1.2529	-0.0538	0.2844
MC47	2	3PLM	0.8099	0.2250	0.1915
MC48	2	3PLM	1.2023	0.3425	0.0264
MC49	2	3PLM	0.7890	-1.0816	0.0900
MC50	2	3PLM	0.9687	1.6943	0.2474
MC51	2	3PLM	0.8166	0.0299	0.2247
MC52	2	3PLM	1.0994	0.3426	0.2816
MC53	2	3PLM	0.8657	0.9133	0.0906
MC54	2	3PLM	0.5647	-0.3270	0.1757
MC55	2	3PLM	0.8781	0.7497	0.2269
MC56	2	3PLM	0.7448	-1.3446	0.2119
MC57	2	3PLM	1.1912	1.1100	0.1533
FR1	3	GPCM	0.5509	0.0414	

Table A7 IRT Item Parameters, Single Group, Design 2, Human Raters, Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.6629	-1.1073	0.1577
MC2	2	3PLM	0.8410	-0.1453	0.1192
MC3	2	3PLM	0.6539	0.5372	0.0808
MC4	2	3PLM	0.7552	1.4968	0.1052
MC5	2	3PLM	1.0160	0.5486	0.2336
MC6	2	3PLM	1.0482	-0.1975	0.2141
MC7	2	3PLM	1.6348	2.3120	0.2249
MC8	2	3PLM	1.1907	-0.4923	0.0465
MC9	2	3PLM	1.0460	-2.2568	0.2408
MC10	2	3PLM	0.9811	-0.5767	0.0987
MC11	2	3PLM	1.3890	1.6572	0.1211
MC12	2	3PLM	1.0612	-1.6597	0.1800
MC13	2	3PLM	1.4239	0.2843	0.1025
MC14	2	3PLM	1.3285	0.5844	0.2560
MC15	2	3PLM	0.8767	-0.7612	0.1425
MC16	2	3PLM	0.8142	-0.5628	0.1864
MC17	2	3PLM	1.0135	0.5887	0.2364
MC18	2	3PLM	1.2266	0.5630	0.2967
MC19	2	3PLM	0.8086	0.6249	0.1574
MC20	2	3PLM	1.0753	-1.3773	0.2059
MC21	2	3PLM	1.1471	0.5885	0.1220
MC22	2	3PLM	0.5675	-0.2249	0.1116
MC23	2	3PLM	1.3862	0.8746	0.2562
MC24	2	3PLM	0.9670	0.3812	0.2055
MC25	2	3PLM	0.7653	0.5840	0.2060
MC26	2	3PLM	1.2059	1.3078	0.2199
MC27	2	3PLM	1.2121	-0.1703	0.1884
MC28	2	3PLM	1.4426	0.5856	0.0533
MC29	2	3PLM	0.9338	-0.6246	0.2493
MC30	2	3PLM	0.8864	2.2825	0.2081
MC31	2	3PLM	1.1335	-1.6562	0.1591
MC32	2	3PLM	0.9325	0.7541	0.0785
MC33	2	3PLM	1.0472	-0.1115	0.2162
MC34	2	3PLM	1.1670	1.1767	0.1683
MC35	2	3PLM	0.7488	2.0440	0.2124
MC36	2	3PLM	1.2311	-1.4391	0.0847
MC37	2	3PLM	0.8883	-0.8425	0.2353
MC38	2	3PLM	1.1661	-0.9064	0.2627
MC39	2	3PLM	1.0059	0.7579	0.2498
MC40	2	3PLM	0.7748	0.5474	0.2519

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	0.9507	-0.4631	0.0597
MC42	2	3PLM	1.4899	0.5516	0.0673
MC43	2	3PLM	0.8333	-2.8551	0.1893
MC44	2	3PLM	0.9926	1.0835	0.1967
MC45	2	3PLM	1.1980	-1.2416	0.1503
MC46	2	3PLM	1.0195	-1.3414	0.1653
MC47	2	3PLM	0.7423	0.1095	0.1292
MC48	2	3PLM	1.0333	0.9283	0.0529
MC49	2	3PLM	0.9025	-0.2822	0.2407
MC50	2	3PLM	0.9565	-0.1655	0.2492
MC51	2	3PLM	0.8160	1.0589	0.0600
MC52	2	3PLM	1.0301	-0.2613	0.1904
MC53	2	3PLM	1.3674	-0.2771	0.2254
MC54	2	3PLM	1.1279	-1.0114	0.1785
FR1	3	GPCM	0.6464	0.2835	
FR2	3	GPCM	0.4200	-0.1340	
FR3	4	GPCM	0.5902	-0.1120	
FR4	4	GPCM	0.6296	-0.1509	
FR5	5	GPCM	0.5642	0.5040	
FR6	5	GPCM	0.5833	-0.5304	

Table A8 Equated IRT Item Parameters, Single Group, Design 2, Human Raters, Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.9592	1.4884	0.0243
MC2	2	3PLM	1.2584	-1.5000	0.1345
MC3	2	3PLM	0.9840	1.6626	0.0390
MC4	2	3PLM	0.8038	-0.2619	0.2560
MC5	2	3PLM	0.9096	0.4619	0.0388
MC6	2	3PLM	1.1206	-0.2489	0.0499
MC7	2	3PLM	1.1591	0.7851	0.0787
MC8	2	3PLM	1.1961	0.5933	0.2370
MC9	2	3PLM	0.8312	0.6683	0.0459
MC10	2	3PLM	1.2112	0.1792	0.0525
MC11	2	3PLM	0.9083	1.1157	0.0320
MC12	2	3PLM	1.1537	0.9034	0.0493
MC13	2	3PLM	0.8609	-0.0212	0.1411
MC14	2	3PLM	0.8797	1.2114	0.2178
MC15	2	3PLM	0.8429	-1.3404	0.3178
MC16	2	3PLM	1.0290	-0.5758	0.1829
MC17	2	3PLM	1.2793	0.8535	0.2754
MC18	2	3PLM	1.1458	-0.6745	0.0376
MC19	2	3PLM	1.1330	-1.8327	0.1995
MC20	2	3PLM	1.0579	0.8262	0.1085
MC21	2	3PLM	1.2026	-0.6937	0.0794
MC22	2	3PLM	0.8836	1.2686	0.1572
MC23	2	3PLM	1.3835	1.5448	0.1607
MC24	2	3PLM	0.8041	0.0197	0.0377
MC25	2	3PLM	1.0526	0.2132	0.0373
MC26	2	3PLM	0.8529	0.3285	0.1139
MC27	2	3PLM	1.1172	1.0347	0.0313
MC28	2	3PLM	0.8246	0.0094	0.2550
MC29	2	3PLM	1.1220	0.6864	0.1505
MC30	2	3PLM	1.3522	1.2414	0.2651
MC31	2	3PLM	0.6534	-0.8010	0.0835
MC32	2	3PLM	0.8427	-0.8304	0.0750
MC33	2	3PLM	1.0009	-0.5336	0.2350
MC34	2	3PLM	0.9418	-1.4611	0.2079
MC35	2	3PLM	1.0020	0.3546	0.0417
MC36	2	3PLM	0.9379	-0.6911	0.1029
MC37	2	3PLM	0.8283	-0.4427	0.3185
MC38	2	3PLM	0.7160	-0.2308	0.2783
MC39	2	3PLM	1.0896	-1.0879	0.1057
MC40	2	3PLM	0.9766	0.7903	0.1878

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	0.8751	2.7207	0.0290
MC42	2	3PLM	1.0341	1.8952	0.1693
MC43	2	3PLM	0.5951	0.0540	0.2232
MC44	2	3PLM	1.0057	-0.3601	0.1881
MC45	2	3PLM	0.8855	-0.5155	0.2407
MC46	2	3PLM	1.0447	-1.9803	0.1434
MC47	2	3PLM	0.8951	0.0216	0.2710
MC48	2	3PLM	0.8227	0.0104	0.2277
MC49	2	3PLM	1.4002	1.1103	0.1540
MC50	2	3PLM	1.0435	0.3570	0.1477
MC51	2	3PLM	0.6265	0.4757	0.0972
MC52	2	3PLM	0.8021	-0.0517	0.1008
MC53	2	3PLM	0.6791	-1.1009	0.2347
MC54	2	3PLM	0.8693	0.6816	0.2442
FR1	3	GPCM	0.4261	0.8815	
FR2	3	GPCM	0.3850	0.7130	
FR3	4	GPCM	0.6815	0.4805	
FR4	4	GPCM	0.6890	-0.2491	
FR5	5	GPCM	0.4419	0.7556	
FR6	5	GPCM	0.6945	-0.8897	

Table A9 IRT Item Parameters, Single Group, Design 1, Automated Raters (Constant Noise), Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0196	-1.4147	0.1020
MC2	2	3PLM	1.0822	1.0118	0.1163
MC3	2	3PLM	1.0377	-1.1823	0.0914
MC4	2	3PLM	1.0401	-1.3869	0.2261
MC5	2	3PLM	1.0318	0.3563	0.0699
MC6	2	3PLM	0.9361	0.8876	0.1621
MC7	2	3PLM	1.3230	-1.0337	0.1703
MC8	2	3PLM	0.7257	0.9473	0.1919
MC9	2	3PLM	0.8842	0.7634	0.1263
MC10	2	3PLM	1.0815	-0.1534	0.0314
MC11	2	3PLM	0.9154	-0.7202	0.2842
MC12	2	3PLM	0.8457	-0.7290	0.1064
MC13	2	3PLM	0.7982	1.3853	0.2557
MC14	2	3PLM	1.0468	-0.8604	0.2641
MC15	2	3PLM	1.0232	-1.9838	0.1358
MC16	2	3PLM	0.5319	-0.5575	0.1925
MC17	2	3PLM	0.8071	1.1954	0.1796
MC18	2	3PLM	0.9606	0.3105	0.0557
MC19	2	3PLM	0.8541	2.4003	0.2852
MC20	2	3PLM	1.0386	0.3389	0.0585
MC21	2	3PLM	1.0364	-0.9806	0.2863
MC22	2	3PLM	0.9984	1.1573	0.0096
MC23	2	3PLM	1.1211	-0.8389	0.1098
MC24	2	3PLM	0.9327	-0.5331	0.0687
MC25	2	3PLM	1.2697	0.4791	0.1310
MC26	2	3PLM	0.8903	-0.4067	0.2987
MC27	2	3PLM	0.8053	0.2225	0.2141
MC28	2	3PLM	1.1289	-0.0184	0.2229
MC29	2	3PLM	1.5044	0.3267	0.0564
MC30	2	3PLM	1.1816	0.0050	0.1264
MC31	2	3PLM	1.3136	-0.0207	0.0200
MC32	2	3PLM	1.1808	-1.5459	0.1422
MC33	2	3PLM	0.9307	0.7995	0.1563
MC34	2	3PLM	1.3154	-0.0467	0.0891
MC35	2	3PLM	1.1547	0.3631	0.1608
MC36	2	3PLM	0.9520	-0.8117	0.0966
MC37	2	3PLM	0.9722	-1.5218	0.2593
MC38	2	3PLM	0.8413	1.6872	0.0376
MC39	2	3PLM	0.9211	-0.5041	0.0680

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	1.1745	0.4869	0.2808
MC41	2	3PLM	0.9754	-0.6350	0.2150
MC42	2	3PLM	0.9729	-0.4889	0.2226
MC43	2	3PLM	0.8817	1.6014	0.2629
MC44	2	3PLM	0.8100	0.3581	0.2846
MC45	2	3PLM	1.0952	0.8969	0.0934
MC46	2	3PLM	1.0532	0.1747	0.2357
MC47	2	3PLM	0.9217	1.4078	0.1784
MC48	2	3PLM	0.8835	0.4095	0.0574
MC49	2	3PLM	1.0074	-0.7639	0.0773
MC50	2	3PLM	0.9635	0.0609	0.2692
MC51	2	3PLM	1.4120	1.5131	0.2421
MC52	2	3PLM	1.6521	-1.2447	0.1416
MC53	2	3PLM	0.9279	0.0535	0.1692
MC54	2	3PLM	1.1015	1.4833	0.3176
MC55	2	3PLM	0.8884	0.3332	0.2767
MC56	2	3PLM	1.0009	-1.2743	0.1467
MC57	2	3PLM	0.9859	0.0859	0.1844
FR1	3	GPCM	0.3920	-2.1209	

Table A10 Equated IRT Item Parameters, Single Group, Design 1, Automated Raters (Constant Noise), Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0478	-1.5235	0.1704
MC2	2	3PLM	1.3813	2.0933	0.0679
MC3	2	3PLM	0.9503	0.3753	0.1491
MC4	2	3PLM	0.9519	-0.1409	0.2206
MC5	2	3PLM	1.1452	0.1142	0.1005
MC6	2	3PLM	1.0467	1.5871	0.3234
MC7	2	3PLM	1.3462	1.7648	0.0227
MC8	2	3PLM	1.0990	-0.3797	0.2060
MC9	2	3PLM	0.8883	-0.8927	0.1504
MC10	2	3PLM	1.4465	0.1068	0.2487
MC11	2	3PLM	1.0201	-0.1281	0.2381
MC12	2	3PLM	0.9974	2.1720	0.2963
MC13	2	3PLM	1.2450	-0.5560	0.1232
MC14	2	3PLM	0.9649	-0.3627	0.1938
MC15	2	3PLM	0.9995	-0.6790	0.2995
MC16	2	3PLM	1.3994	0.0415	0.2656
MC17	2	3PLM	1.9323	1.7073	0.2423
MC18	2	3PLM	1.2895	-0.1174	0.0378
MC19	2	3PLM	1.1752	-1.1011	0.2641
MC20	2	3PLM	1.1192	-0.4151	0.2055
MC21	2	3PLM	1.0999	0.0038	0.1766
MC22	2	3PLM	1.1649	0.8157	0.1780
MC23	2	3PLM	1.2133	-0.5801	0.2268
MC24	2	3PLM	0.9877	0.4845	0.2577
MC25	2	3PLM	1.1095	0.6486	0.0449
MC26	2	3PLM	0.9519	-1.2858	0.2442
MC27	2	3PLM	0.9805	0.8788	0.2380
MC28	2	3PLM	1.1011	1.0288	0.1186
MC29	2	3PLM	1.3168	1.1399	0.0834
MC30	2	3PLM	0.9777	0.8249	0.1344
MC31	2	3PLM	1.2062	-0.4717	0.2458
MC32	2	3PLM	0.9150	0.2767	0.0692
MC33	2	3PLM	1.2816	1.4183	0.0496
MC34	2	3PLM	0.9614	-0.0082	0.1845
MC35	2	3PLM	0.9247	-0.1648	0.2676
MC36	2	3PLM	1.5313	0.0467	0.1720
MC37	2	3PLM	1.2250	-0.6110	0.0584
MC38	2	3PLM	0.9182	-0.1386	0.1488
MC39	2	3PLM	1.1706	-1.0964	0.2914

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	1.0653	-0.3997	0.3122
MC41	2	3PLM	0.9157	-1.1348	0.2104
MC42	2	3PLM	0.8539	-1.4155	0.1455
MC43	2	3PLM	0.9659	0.0707	0.1151
MC44	2	3PLM	1.2420	0.2783	0.0907
MC45	2	3PLM	0.7868	-2.3294	0.2202
MC46	2	3PLM	1.2651	-0.0215	0.2877
MC47	2	3PLM	0.8015	0.2439	0.1883
MC48	2	3PLM	1.2111	0.3689	0.0278
MC49	2	3PLM	0.7866	-1.0380	0.1089
MC50	2	3PLM	0.9889	1.7101	0.2490
MC51	2	3PLM	0.8387	0.0785	0.2359
MC52	2	3PLM	1.1214	0.3745	0.2850
MC53	2	3PLM	0.8876	0.9347	0.0942
MC54	2	3PLM	0.5784	-0.2338	0.2012
MC55	2	3PLM	0.8877	0.7744	0.2294
MC56	2	3PLM	0.7439	-1.2921	0.2348
MC57	2	3PLM	1.2223	1.1254	0.1553
FR1	3	GPCM	0.5605	0.2597	

Table A11 IRT Item Parameters, Single Group, Design 2, Automated Raters (Constant Noise), Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.6696	-1.0496	0.1839
MC2	2	3PLM	0.8583	-0.1186	0.1276
MC3	2	3PLM	0.6837	0.5739	0.0956
MC4	2	3PLM	0.7936	1.4800	0.1102
MC5	2	3PLM	1.0467	0.5635	0.2384
MC6	2	3PLM	1.0704	-0.1646	0.2246
MC7	2	3PLM	1.7230	2.2438	0.2249
MC8	2	3PLM	1.2019	-0.4660	0.0575
MC9	2	3PLM	1.0104	-2.3208	0.2347
MC10	2	3PLM	0.9915	-0.5492	0.1097
MC11	2	3PLM	1.4163	1.6287	0.1205
MC12	2	3PLM	1.0365	-1.6733	0.1895
MC13	2	3PLM	1.4537	0.2985	0.1050
MC14	2	3PLM	1.3668	0.5911	0.2576
MC15	2	3PLM	0.8820	-0.7313	0.1558
MC16	2	3PLM	0.8330	-0.5153	0.2040
MC17	2	3PLM	1.0462	0.5972	0.2397
MC18	2	3PLM	1.2589	0.5675	0.2975
MC19	2	3PLM	0.8308	0.6355	0.1620
MC20	2	3PLM	1.0631	-1.3619	0.2254
MC21	2	3PLM	1.1977	0.6003	0.1276
MC22	2	3PLM	0.5782	-0.1791	0.1264
MC23	2	3PLM	1.4455	0.8740	0.2584
MC24	2	3PLM	1.0051	0.4044	0.2133
MC25	2	3PLM	0.7944	0.6117	0.2159
MC26	2	3PLM	1.2807	1.2864	0.2222
MC27	2	3PLM	1.2163	-0.1566	0.1898
MC28	2	3PLM	1.4818	0.5930	0.0552
MC29	2	3PLM	0.9541	-0.5760	0.2681
MC30	2	3PLM	0.9091	2.2333	0.2078
MC31	2	3PLM	1.1158	-1.6456	0.1869
MC32	2	3PLM	0.9559	0.7541	0.0800
MC33	2	3PLM	1.0732	-0.0791	0.2263
MC34	2	3PLM	1.2040	1.1648	0.1692
MC35	2	3PLM	0.7916	1.9947	0.2150
MC36	2	3PLM	1.2048	-1.4472	0.0928
MC37	2	3PLM	0.8982	-0.8076	0.2495
MC38	2	3PLM	1.1823	-0.8649	0.2836
MC39	2	3PLM	1.0344	0.7605	0.2518

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	0.8036	0.5634	0.2579
MC41	2	3PLM	0.9626	-0.4343	0.0710
MC42	2	3PLM	1.5347	0.5581	0.0687
MC43	2	3PLM	0.8007	-2.9359	0.1997
MC44	2	3PLM	1.0275	1.0760	0.1987
MC45	2	3PLM	1.1827	-1.2213	0.1734
MC46	2	3PLM	1.0158	-1.3086	0.1941
MC47	2	3PLM	0.7686	0.1505	0.1437
MC48	2	3PLM	1.0603	0.9245	0.0542
MC49	2	3PLM	0.9246	-0.2483	0.2512
MC50	2	3PLM	0.9783	-0.1325	0.2590
MC51	2	3PLM	0.8407	1.0578	0.0636
MC52	2	3PLM	1.0433	-0.2360	0.1978
MC53	2	3PLM	1.3722	-0.2583	0.2291
MC54	2	3PLM	1.1325	-0.9767	0.2007
FR1	3	GPCM	0.6707	0.4140	
FR2	3	GPCM	0.4147	0.0754	
FR3	4	GPCM	0.6816	0.0092	
FR4	5	GPCM	0.5018	-0.3472	

Table A12 Equated IRT Item Parameters, Single Group, Design 2, Automated Raters (Constant Noise), Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.9993	1.4613	0.0256
MC2	2	3PLM	1.2343	-1.5004	0.1506
MC3	2	3PLM	1.0135	1.6348	0.0394
MC4	2	3PLM	0.8214	-0.2185	0.2698
MC5	2	3PLM	0.9271	0.4741	0.0426
MC6	2	3PLM	1.1272	-0.2300	0.0551
MC7	2	3PLM	1.1902	0.7844	0.0797
MC8	2	3PLM	1.2455	0.6059	0.2417
MC9	2	3PLM	0.8498	0.6741	0.0489
MC10	2	3PLM	1.2319	0.1970	0.0568
MC11	2	3PLM	0.9410	1.1064	0.0345
MC12	2	3PLM	1.1961	0.8987	0.0510
MC13	2	3PLM	0.8680	-0.0025	0.1459
MC14	2	3PLM	0.9272	1.2037	0.2226
MC15	2	3PLM	0.8377	-1.3270	0.3278
MC16	2	3PLM	1.0455	-0.5373	0.1987
MC17	2	3PLM	1.3186	0.8504	0.2760
MC18	2	3PLM	1.1540	-0.6475	0.0511
MC19	2	3PLM	1.1007	-1.8552	0.2116
MC20	2	3PLM	1.0933	0.8262	0.1109
MC21	2	3PLM	1.2088	-0.6715	0.0895
MC22	2	3PLM	0.9361	1.2554	0.1622
MC23	2	3PLM	1.4488	1.5141	0.1612
MC24	2	3PLM	0.8126	0.0344	0.0414
MC25	2	3PLM	1.0752	0.2284	0.0414
MC26	2	3PLM	0.8748	0.3475	0.1205
MC27	2	3PLM	1.1640	1.0264	0.0335
MC28	2	3PLM	0.8492	0.0481	0.2670
MC29	2	3PLM	1.1649	0.6905	0.1534
MC30	2	3PLM	1.4080	1.2235	0.2656
MC31	2	3PLM	0.6554	-0.7743	0.0946
MC32	2	3PLM	0.8498	-0.7976	0.0909
MC33	2	3PLM	1.0110	-0.5082	0.2432
MC34	2	3PLM	0.9310	-1.4501	0.2241
MC35	2	3PLM	1.0228	0.3640	0.0439
MC36	2	3PLM	0.9406	-0.6721	0.1105
MC37	2	3PLM	0.8491	-0.3913	0.3346
MC38	2	3PLM	0.7386	-0.1823	0.2926
MC39	2	3PLM	1.0752	-1.0817	0.1133

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	1.0118	0.7960	0.1918
MC41	2	3PLM	0.9012	2.6666	0.0294
MC42	2	3PLM	1.0614	1.8606	0.1694
MC43	2	3PLM	0.6280	0.1319	0.2470
MC44	2	3PLM	1.0227	-0.3313	0.1974
MC45	2	3PLM	0.9057	-0.4655	0.2593
MC46	2	3PLM	1.0135	-2.0219	0.1440
MC47	2	3PLM	0.9163	0.0475	0.2777
MC48	2	3PLM	0.8428	0.0359	0.2349
MC49	2	3PLM	1.4447	1.0959	0.1536
MC50	2	3PLM	1.0746	0.3741	0.1529
MC51	2	3PLM	0.6456	0.4954	0.1049
MC52	2	3PLM	0.8221	-0.0171	0.1130
MC53	2	3PLM	0.6752	-1.0904	0.2406
MC54	2	3PLM	0.9063	0.6956	0.2507
FR1	3	GPCM	0.4353	1.1433	
FR2	3	GPCM	0.3944	0.9393	
FR3	4	GPCM	0.7812	-0.0891	
FR4	5	GPCM	0.5795	-0.7630	

Table A13 IRT Item Parameters, Single Group, Design 1, Automated Raters (Reduced Noise), Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0308	-1.4269	0.0933
MC2	2	3PLM	1.0699	0.9958	0.1153
MC3	2	3PLM	1.0489	-1.1938	0.0871
MC4	2	3PLM	1.0474	-1.3926	0.2259
MC5	2	3PLM	1.0223	0.3320	0.0666
MC6	2	3PLM	0.9006	0.8675	0.1554
MC7	2	3PLM	1.3309	-1.0428	0.1723
MC8	2	3PLM	0.7352	0.9327	0.1955
MC9	2	3PLM	0.8966	0.7417	0.1271
MC10	2	3PLM	1.0803	-0.1749	0.0290
MC11	2	3PLM	0.9237	-0.7195	0.2920
MC12	2	3PLM	0.8555	-0.7365	0.1091
MC13	2	3PLM	0.8024	1.3596	0.2560
MC14	2	3PLM	1.0538	-0.8755	0.2616
MC15	2	3PLM	1.0277	-1.9726	0.1531
MC16	2	3PLM	0.5352	-0.5710	0.1947
MC17	2	3PLM	0.8157	1.1665	0.1799
MC18	2	3PLM	0.9634	0.2897	0.0548
MC19	2	3PLM	0.8388	2.4053	0.2849
MC20	2	3PLM	1.0469	0.3209	0.0593
MC21	2	3PLM	1.0531	-0.9834	0.2895
MC22	2	3PLM	0.9949	1.1409	0.0093
MC23	2	3PLM	1.1257	-0.8549	0.1067
MC24	2	3PLM	0.9391	-0.5500	0.0651
MC25	2	3PLM	1.2685	0.4636	0.1324
MC26	2	3PLM	0.8975	-0.4208	0.2985
MC27	2	3PLM	0.7944	0.1927	0.2095
MC28	2	3PLM	1.1284	-0.0459	0.2201
MC29	2	3PLM	1.5001	0.3070	0.0558
MC30	2	3PLM	1.1935	-0.0093	0.1282
MC31	2	3PLM	1.3186	-0.0388	0.0199
MC32	2	3PLM	1.1854	-1.5558	0.1416
MC33	2	3PLM	0.9251	0.7742	0.1533
MC34	2	3PLM	1.3142	-0.0687	0.0866
MC35	2	3PLM	1.1714	0.3488	0.1641
MC36	2	3PLM	0.9551	-0.8323	0.0935
MC37	2	3PLM	0.9697	-1.5638	0.2379
MC38	2	3PLM	0.8252	1.6801	0.0355
MC39	2	3PLM	0.9226	-0.5240	0.0649

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	1.1843	0.4620	0.2794
MC41	2	3PLM	0.9662	-0.6742	0.2005
MC42	2	3PLM	0.9617	-0.5216	0.2139
MC43	2	3PLM	0.8610	1.5925	0.2618
MC44	2	3PLM	0.8159	0.3448	0.2862
MC45	2	3PLM	1.0781	0.8819	0.0918
MC46	2	3PLM	1.0586	0.1586	0.2363
MC47	2	3PLM	0.9022	1.4015	0.1760
MC48	2	3PLM	0.8920	0.3899	0.0578
MC49	2	3PLM	1.0118	-0.7803	0.0751
MC50	2	3PLM	0.9501	0.0321	0.2644
MC51	2	3PLM	1.4239	1.4933	0.2426
MC52	2	3PLM	1.6546	-1.2662	0.1319
MC53	2	3PLM	0.9250	0.0412	0.1713
MC54	2	3PLM	1.0684	1.4813	0.3163
MC55	2	3PLM	0.8987	0.3125	0.2767
MC56	2	3PLM	1.0195	-1.2678	0.1541
MC57	2	3PLM	0.9786	0.0691	0.1844
FR1	3	GPCM	0.3975	0.7130	

Table A14 Equated IRT Item Parameters, Single Group, Design 1, Automated Raters (Reduced Noise), Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0575	-1.5381	0.1604
MC2	2	3PLM	1.3234	2.1201	0.0679
MC3	2	3PLM	0.9502	0.3518	0.1488
MC4	2	3PLM	0.9479	-0.1708	0.2152
MC5	2	3PLM	1.1279	0.0831	0.0934
MC6	2	3PLM	1.0238	1.5673	0.3210
MC7	2	3PLM	1.3062	1.7695	0.0219
MC8	2	3PLM	1.0941	-0.4040	0.2021
MC9	2	3PLM	0.8802	-0.9272	0.1401
MC10	2	3PLM	1.4500	0.0900	0.2495
MC11	2	3PLM	1.0295	-0.1392	0.2412
MC12	2	3PLM	0.9814	2.1589	0.2958
MC13	2	3PLM	1.2395	-0.5826	0.1160
MC14	2	3PLM	0.9660	-0.3803	0.1923
MC15	2	3PLM	1.0000	-0.6975	0.2992
MC16	2	3PLM	1.4126	0.0247	0.2668
MC17	2	3PLM	1.8912	1.7665	0.2422
MC18	2	3PLM	1.2962	-0.1390	0.0365
MC19	2	3PLM	1.1911	-1.1069	0.2665
MC20	2	3PLM	1.1273	-0.4342	0.2027
MC21	2	3PLM	1.0984	-0.0149	0.1771
MC22	2	3PLM	1.1771	0.7941	0.1781
MC23	2	3PLM	1.2261	-0.5873	0.2316
MC24	2	3PLM	1.0013	0.4752	0.2624
MC25	2	3PLM	1.1108	0.6282	0.0442
MC26	2	3PLM	0.9626	-1.2967	0.2411
MC27	2	3PLM	0.9715	0.8520	0.2352
MC28	2	3PLM	1.0917	1.0078	0.1168
MC29	2	3PLM	1.2952	1.1222	0.0819
MC30	2	3PLM	1.0007	0.8011	0.1365
MC31	2	3PLM	1.2072	-0.4918	0.2436
MC32	2	3PLM	0.9113	0.2559	0.0673
MC33	2	3PLM	1.2704	1.4056	0.0488
MC34	2	3PLM	0.9653	-0.0256	0.1842
MC35	2	3PLM	0.9223	-0.1923	0.2629
MC36	2	3PLM	1.5351	0.0286	0.1718
MC37	2	3PLM	1.2383	-0.6194	0.0619
MC38	2	3PLM	0.9126	-0.1681	0.1432
MC39	2	3PLM	1.1740	-1.1114	0.2900

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	1.0654	-0.4242	0.3076
MC41	2	3PLM	0.9158	-1.1607	0.2014
MC42	2	3PLM	0.8580	-1.4338	0.1362
MC43	2	3PLM	0.9773	0.0528	0.1160
MC44	2	3PLM	1.2512	0.2608	0.0910
MC45	2	3PLM	0.7960	-2.3320	0.2115
MC46	2	3PLM	1.2656	-0.0400	0.2874
MC47	2	3PLM	0.8216	0.2422	0.1967
MC48	2	3PLM	1.2109	0.3491	0.0272
MC49	2	3PLM	0.7906	-1.0648	0.0989
MC50	2	3PLM	0.9978	1.6781	0.2486
MC51	2	3PLM	0.8260	0.0483	0.2300
MC52	2	3PLM	1.1124	0.3541	0.2842
MC53	2	3PLM	0.8858	0.9156	0.0934
MC54	2	3PLM	0.5748	-0.2832	0.1891
MC55	2	3PLM	0.8917	0.7546	0.2288
MC56	2	3PLM	0.7467	-1.3156	0.2273
MC57	2	3PLM	1.2154	1.1088	0.1547
FR1	3	GPCM	0.8044	0.5374	

Table A15 IRT Item Parameters, Single Group, Design 2, Automated Raters (Reduced Noise), Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.6652	-1.0972	0.1610
MC2	2	3PLM	0.8407	-0.1458	0.1176
MC3	2	3PLM	0.6602	0.5427	0.0830
MC4	2	3PLM	0.7696	1.4883	0.1070
MC5	2	3PLM	1.0263	0.5536	0.2351
MC6	2	3PLM	1.0490	-0.1926	0.2149
MC7	2	3PLM	1.6595	2.2972	0.2251
MC8	2	3PLM	1.1942	-0.4842	0.0499
MC9	2	3PLM	1.0356	-2.2676	0.2444
MC10	2	3PLM	0.9831	-0.5711	0.1002
MC11	2	3PLM	1.4052	1.6479	0.1213
MC12	2	3PLM	1.0503	-1.6706	0.1778
MC13	2	3PLM	1.4330	0.2890	0.1036
MC14	2	3PLM	1.3446	0.5857	0.2566
MC15	2	3PLM	0.8793	-0.7554	0.1441
MC16	2	3PLM	0.8201	-0.5465	0.1923
MC17	2	3PLM	1.0321	0.5947	0.2394
MC18	2	3PLM	1.2358	0.5615	0.2961
MC19	2	3PLM	0.8134	0.6245	0.1576
MC20	2	3PLM	1.0667	-1.3872	0.2007
MC21	2	3PLM	1.1636	0.5912	0.1235
MC22	2	3PLM	0.5730	-0.2071	0.1172
MC23	2	3PLM	1.4117	0.8760	0.2575
MC24	2	3PLM	0.9769	0.3835	0.2061
MC25	2	3PLM	0.7710	0.5890	0.2076
MC26	2	3PLM	1.2276	1.2983	0.2203
MC27	2	3PLM	1.2207	-0.1598	0.1920
MC28	2	3PLM	1.4478	0.5867	0.0533
MC29	2	3PLM	0.9372	-0.6148	0.2528
MC30	2	3PLM	0.8944	2.2732	0.2085
MC31	2	3PLM	1.1300	-1.6408	0.1773
MC32	2	3PLM	0.9380	0.7532	0.0786
MC33	2	3PLM	1.0588	-0.1004	0.2196
MC34	2	3PLM	1.1703	1.1739	0.1681
MC35	2	3PLM	0.7512	2.0319	0.2119
MC36	2	3PLM	1.2259	-1.4394	0.0860
MC37	2	3PLM	0.8870	-0.8436	0.2338
MC38	2	3PLM	1.1807	-0.8791	0.2770
MC39	2	3PLM	1.0104	0.7546	0.2490

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	0.7813	0.5467	0.2521
MC41	2	3PLM	0.9512	-0.4571	0.0618
MC42	2	3PLM	1.5062	0.5535	0.0681
MC43	2	3PLM	0.8236	-2.8707	0.1979
MC44	2	3PLM	1.0061	1.0810	0.1977
MC45	2	3PLM	1.1890	-1.2406	0.1536
MC46	2	3PLM	1.0188	-1.3360	0.1695
MC47	2	3PLM	0.7473	0.1204	0.1328
MC48	2	3PLM	1.0388	0.9274	0.0532
MC49	2	3PLM	0.9077	-0.2788	0.2408
MC50	2	3PLM	0.9607	-0.1579	0.2511
MC51	2	3PLM	0.8217	1.0573	0.0606
MC52	2	3PLM	1.0380	-0.2542	0.1920
MC53	2	3PLM	1.3762	-0.2690	0.2276
MC54	2	3PLM	1.1293	-1.0060	0.1810
FR1	3	GPCM	0.6550	0.4149	
FR2	3	GPCM	0.4007	0.0976	
FR3	4	GPCM	0.6605	0.0688	
FR4	4	GPCM	0.6953	0.0153	
FR5	5	GPCM	0.5173	0.6845	
FR6	5	GPCM	0.5204	-0.3609	

Table A16 Equated IRT Item Parameters, Single Group, Design 2, Automated Raters (Reduced Noise), Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.9644	1.4809	0.0240
MC2	2	3PLM	1.2565	-1.5027	0.1327
MC3	2	3PLM	0.9946	1.6528	0.0392
MC4	2	3PLM	0.8005	-0.2696	0.2520
MC5	2	3PLM	0.9140	0.4673	0.0405
MC6	2	3PLM	1.1186	-0.2467	0.0495
MC7	2	3PLM	1.1739	0.7859	0.0799
MC8	2	3PLM	1.2034	0.5966	0.2379
MC9	2	3PLM	0.8374	0.6690	0.0467
MC10	2	3PLM	1.2181	0.1831	0.0530
MC11	2	3PLM	0.9187	1.1143	0.0334
MC12	2	3PLM	1.1699	0.9021	0.0503
MC13	2	3PLM	0.8680	-0.0087	0.1457
MC14	2	3PLM	0.8930	1.2138	0.2200
MC15	2	3PLM	0.8418	-1.3460	0.3143
MC16	2	3PLM	1.0396	-0.5590	0.1899
MC17	2	3PLM	1.3071	0.8556	0.2773
MC18	2	3PLM	1.1478	-0.6679	0.0401
MC19	2	3PLM	1.1225	-1.8514	0.1898
MC20	2	3PLM	1.0734	0.8267	0.1099
MC21	2	3PLM	1.2068	-0.6866	0.0820
MC22	2	3PLM	0.9051	1.2655	0.1598
MC23	2	3PLM	1.3888	1.5391	0.1606
MC24	2	3PLM	0.8091	0.0259	0.0395
MC25	2	3PLM	1.0594	0.2180	0.0388
MC26	2	3PLM	0.8585	0.3334	0.1156
MC27	2	3PLM	1.1291	1.0310	0.0315
MC28	2	3PLM	0.8328	0.0170	0.2572
MC29	2	3PLM	1.1269	0.6857	0.1501
MC30	2	3PLM	1.3843	1.2387	0.2665
MC31	2	3PLM	0.6543	-0.7971	0.0843
MC32	2	3PLM	0.8472	-0.8153	0.0821
MC33	2	3PLM	1.0123	-0.5172	0.2411
MC34	2	3PLM	0.9421	-1.4464	0.2187
MC35	2	3PLM	1.0131	0.3585	0.0432
MC36	2	3PLM	0.9387	-0.6861	0.1044
MC37	2	3PLM	0.8340	-0.4305	0.3219
MC38	2	3PLM	0.7196	-0.2232	0.2801
MC39	2	3PLM	1.0902	-1.0756	0.1146

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	0.9834	0.7931	0.1890
MC41	2	3PLM	0.8842	2.7097	0.0293
MC42	2	3PLM	1.0342	1.8836	0.1685
MC43	2	3PLM	0.6015	0.0755	0.2295
MC44	2	3PLM	1.0130	-0.3469	0.1933
MC45	2	3PLM	0.8910	-0.5057	0.2437
MC46	2	3PLM	1.0333	-1.9924	0.1437
MC47	2	3PLM	0.9050	0.0335	0.2745
MC48	2	3PLM	0.8231	0.0102	0.2269
MC49	2	3PLM	1.4071	1.1093	0.1542
MC50	2	3PLM	1.0470	0.3591	0.1478
MC51	2	3PLM	0.6286	0.4813	0.0988
MC52	2	3PLM	0.8123	-0.0368	0.1063
MC53	2	3PLM	0.6855	-1.0722	0.2451
MC54	2	3PLM	0.8718	0.6803	0.2440
FR1	3	GPCM	0.4254	0.8840	
FR2	3	GPCM	0.3855	0.7124	
FR3	4	GPCM	0.6855	0.4813	
FR4	4	GPCM	0.6905	-0.2464	
FR5	5	GPCM	0.4427	0.7551	
FR6	5	GPCM	0.6948	-0.8893	

Table A17 IRT Item Parameters, NEAT, Design 1, Ideal Raters, Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.9907	0.2619	0.2080
MC2	2	3PLM	0.8366	-0.5267	0.1954
MC3	2	3PLM	1.0187	0.2100	0.0590
MC4	2	3PLM	0.8489	1.0085	0.2289
MC5	2	3PLM	0.8389	0.3596	0.2827
MC6	2	3PLM	0.8430	0.5162	0.1961
MC7	2	3PLM	1.0849	0.6417	0.1690
MC8	2	3PLM	1.5616	1.0445	0.1922
MC9	2	3PLM	1.3577	0.1792	0.0425
MC10	2	3PLM	1.1355	-0.2969	0.2848
MC11	2	3PLM	1.0702	0.0747	0.0483
MC12	2	3PLM	1.5008	-0.1742	0.0239
MC13	2	3PLM	1.2664	1.1485	0.2648
MC14	2	3PLM	1.3577	1.0226	0.2973
MC15	2	3PLM	0.9455	-2.3509	0.2356
MC16	2	3PLM	1.1228	-0.7643	0.0516
MC17	2	3PLM	0.7503	-0.9864	0.1136
MC18	2	3PLM	0.8372	0.0426	0.1238
MC19	2	3PLM	1.2393	-0.3382	0.1773
MC20	2	3PLM	1.1886	0.2386	0.0985
MC21	2	3PLM	0.7100	-0.3540	0.1057
MC22	2	3PLM	1.0558	0.2788	0.2283
MC23	2	3PLM	1.1569	-1.0024	0.0834
MC24	2	3PLM	1.1049	-0.5064	0.1755
MC25	2	3PLM	1.4797	-0.7115	0.0781
MC26	2	3PLM	0.8816	0.1759	0.0760
MC27	2	3PLM	0.7693	-0.0203	0.1690
MC28	2	3PLM	1.0174	0.2564	0.1372
MC29	2	3PLM	0.9447	-0.7037	0.1446
MC30	2	3PLM	0.9802	0.1614	0.2408
MC31	2	3PLM	0.9814	-1.9915	0.1676
MC32	2	3PLM	0.9882	-0.0459	0.0299
MC33	2	3PLM	0.9411	0.8308	0.2501
MC34	2	3PLM	1.3708	-0.2292	0.0605
MC35	2	3PLM	1.1295	0.5202	0.2337
MC36	2	3PLM	1.0385	0.0817	0.0594
MC37	2	3PLM	0.8795	0.4348	0.1401
MC38	2	3PLM	0.8706	-0.6907	0.0844
MC39	2	3PLM	1.5381	-0.9949	0.1718
MC40	2	3PLM	1.3249	-1.1584	0.3094

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.0602	-1.0368	0.0977
MC42	2	3PLM	1.5464	-0.5733	0.2681
MC43	2	3PLM	1.0629	-0.7659	0.1328
MC44	2	3PLM	1.1470	1.4509	0.2226
MC45	2	3PLM	1.3127	0.4658	0.2876
MC46	2	3PLM	0.9301	-0.5564	0.1797
MC47	2	3PLM	1.0168	0.5473	0.2388
MC48	2	3PLM	1.1907	-0.8840	0.0000
MC49	2	3PLM	0.9398	-1.1147	0.2326
MC50	2	3PLM	0.9538	-0.1020	0.2935
MC51	2	3PLM	0.8904	-0.8661	0.1232
MC52	2	3PLM	1.0058	0.1382	0.3282
MC53	2	3PLM	0.6657	1.9311	0.2968
MC54	2	3PLM	0.8833	-1.5983	0.2933
MC55	2	3PLM	0.9425	0.8659	0.1381
MC56	2	3PLM	1.4081	0.0330	0.1019
MC57	2	3PLM	0.8911	-0.4814	0.3095
FR1	3	GPCM	0.5975	-1.2277	
FR2	4	GPCM	1.0705	-0.8841	
FR3	5	GPCM	1.1602	-0.0569	

Table A18 Equated IRT Item Parameters, NEAT, Design 1, Ideal Raters, Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.9640	2.4776	0.0768
MC2	2	3PLM	0.8962	-0.2274	0.1373
MC3	2	3PLM	1.2283	1.4928	0.2634
MC4	2	3PLM	1.1456	0.3052	0.1662
MC5	2	3PLM	1.3061	0.0017	0.0832
MC6	2	3PLM	1.0612	0.2151	0.2003
MC7	2	3PLM	1.0282	1.3575	0.0914
MC8	2	3PLM	0.8489	-1.4659	0.1633
MC9	2	3PLM	0.9283	-1.1032	0.1162
MC10	2	3PLM	0.9088	0.2475	0.1013
MC11	2	3PLM	1.6919	1.3590	0.0551
MC12	2	3PLM	1.1622	0.3441	0.0338
MC13	2	3PLM	1.1238	0.5504	0.0681
MC14	2	3PLM	1.0992	-0.3678	0.1245
MC15	2	3PLM	0.8728	-0.9560	0.1683
MC16	2	3PLM	0.8543	0.4126	0.1067
MC17	2	3PLM	0.9084	0.3451	0.0791
MC18	2	3PLM	1.0281	-1.3096	0.1616
MC19	2	3PLM	0.2771	-0.1300	0.0997
MC20	2	3PLM	0.8223	1.7750	0.3064
MC21	2	3PLM	1.1005	0.6588	0.0957
MC22	2	3PLM	0.9388	0.4320	0.1954
MC23	2	3PLM	1.2221	-1.3609	0.2001
MC24	2	3PLM	1.3308	-0.4527	0.0575
MC25	2	3PLM	1.1460	1.8166	0.2996
MC26	2	3PLM	0.8770	-0.0393	0.0534
MC27	2	3PLM	0.8786	1.7648	0.0434
MC28	2	3PLM	1.0410	-0.5091	0.2731
MC29	2	3PLM	1.0730	-0.5701	0.2280
MC30	2	3PLM	0.9128	-0.8273	0.1115
MC31	2	3PLM	0.9138	-2.7649	0.2232
MC32	2	3PLM	1.0824	-1.4382	0.1959
MC33	2	3PLM	1.3603	0.1702	0.0364
MC34	2	3PLM	1.1783	-1.3013	0.1561
MC35	2	3PLM	0.9629	0.7313	0.1238
MC36	2	3PLM	1.1833	-0.9413	0.0859
MC37	2	3PLM	0.9971	0.3068	0.2252
MC38	2	3PLM	1.0712	0.2187	0.0689
MC39	2	3PLM	1.1056	0.6662	0.1639
MC40	2	3PLM	1.3436	0.1836	0.0446

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.1736	0.0912	0.0729
MC42	2	3PLM	1.3756	1.1097	0.2472
MC43	2	3PLM	0.8951	-2.5597	0.1792
MC44	2	3PLM	1.2113	-0.3637	0.1598
MC45	2	3PLM	0.6826	-0.3432	0.1085
MC46	2	3PLM	1.5242	-0.7031	0.0621
MC47	2	3PLM	0.7365	-0.1497	0.1193
MC48	2	3PLM	1.0590	-1.8996	0.1863
MC49	2	3PLM	0.9704	0.8572	0.2548
MC50	2	3PLM	0.9793	0.4663	0.1740
MC51	2	3PLM	1.5515	-0.9016	0.2332
MC52	2	3PLM	1.2059	-0.6159	0.2183
MC53	2	3PLM	1.0033	0.4147	0.2568
MC54	2	3PLM	1.1384	0.5531	0.2550
MC55	2	3PLM	0.9100	-0.8386	0.1521
MC56	2	3PLM	0.7462	1.8505	0.3038
MC57	2	3PLM	0.9033	-0.5247	0.2873
FR1	3	GPCM	0.6295	-1.1977	
FR2	4	GPCM	1.1056	-0.8779	
FR3	5	GPCM	1.1106	-0.0507	

Table A19 IRT Item Parameters, NEAT, Design 2, Ideal Raters, Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.6828	0.3543	0.1898
MC2	2	3PLM	0.8473	-1.6323	0.2222
MC3	2	3PLM	0.8693	-0.6672	0.2688
MC4	2	3PLM	0.8266	-0.6558	0.1123
MC5	2	3PLM	1.2228	0.6700	0.0918
MC6	2	3PLM	0.9904	-1.2468	0.1917
MC7	2	3PLM	1.0627	0.6725	0.2467
MC8	2	3PLM	0.8358	-0.4989	0.3150
MC9	2	3PLM	1.3578	0.5189	0.2756
MC10	2	3PLM	0.7797	2.3551	0.1659
MC11	2	3PLM	1.1667	-0.7000	0.2164
MC12	2	3PLM	1.1218	1.1801	0.1778
MC13	2	3PLM	1.0996	-1.0665	0.2119
MC14	2	3PLM	0.7766	-0.9984	0.1086
MC15	2	3PLM	1.5677	0.2445	0.0450
MC16	2	3PLM	0.9032	-2.0058	0.0000
MC17	2	3PLM	0.9382	-2.8916	0.2131
MC18	2	3PLM	0.9230	-1.0557	0.0820
MC19	2	3PLM	0.6301	1.0378	0.0805
MC20	2	3PLM	0.8277	-0.4821	0.0931
MC21	2	3PLM	0.7758	-1.6785	0.0000
MC22	2	3PLM	0.8364	0.2363	0.2346
MC23	2	3PLM	1.0455	1.0571	0.1968
MC24	2	3PLM	1.0665	-0.2539	0.0546
MC25	2	3PLM	1.0372	-1.5281	0.0000
MC26	2	3PLM	0.7744	0.6752	0.0867
MC27	2	3PLM	1.1438	0.1964	0.2188
MC28	2	3PLM	0.8474	-1.5438	0.1508
MC29	2	3PLM	1.2431	-0.1262	0.0292
MC30	2	3PLM	1.2139	0.9358	0.0135
MC31	2	3PLM	0.9729	0.0425	0.1090
MC32	2	3PLM	0.8494	-0.4406	0.1800
MC33	2	3PLM	0.9212	-1.9896	0.1498
MC34	2	3PLM	1.1113	0.7216	0.2313
MC35	2	3PLM	1.0574	-1.7430	0.2169
MC36	2	3PLM	1.1102	0.8835	0.1867
MC37	2	3PLM	1.0221	1.0084	0.0935
MC38	2	3PLM	1.1657	-0.2484	0.0418
MC39	2	3PLM	0.9616	-0.2717	0.2103
MC40	2	3PLM	0.8647	-0.0222	0.1517

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.4324	1.3370	0.0235
MC42	2	3PLM	1.2038	-0.7144	0.0510
MC43	2	3PLM	1.0423	0.9690	0.0841
MC44	2	3PLM	0.7716	-0.6333	0.0628
MC45	2	3PLM	1.1904	1.2389	0.0093
MC46	2	3PLM	0.8334	-0.0791	0.3020
MC47	2	3PLM	0.8860	0.8493	0.2066
MC48	2	3PLM	1.2055	-1.1972	0.1590
MC49	2	3PLM	1.4199	0.5787	0.0347
MC50	2	3PLM	1.0939	-0.2031	0.2570
MC51	2	3PLM	1.0629	-0.4711	0.2270
MC52	2	3PLM	0.7671	-1.6897	0.1492
MC53	2	3PLM	1.3676	0.4046	0.0920
MC54	2	3PLM	0.7536	1.4102	0.1562
FR1	3	GPCM	1.3668	0.9946	
FR2	3	GPCM	0.8923	-0.3398	
FR3	4	GPCM	1.4991	-0.2279	
FR4	4	GPCM	1.1819	-0.7758	
FR5	5	GPCM	0.7344	0.5574	
FR6	5	GPCM	0.9514	0.2564	

Table A20 Equated IRT Item Parameters, NEAT, Design 2, Ideal Raters, Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.2330	-1.4031	0.1955
MC2	2	3PLM	0.9016	1.6929	0.2899
MC3	2	3PLM	0.7402	-1.0682	0.2368
MC4	2	3PLM	0.8158	-0.4747	0.2053
MC5	2	3PLM	1.2475	-1.0077	0.1240
MC6	2	3PLM	1.3482	0.7657	0.2617
MC7	2	3PLM	1.0647	-0.3517	0.0377
MC8	2	3PLM	0.9015	-0.9525	0.2294
MC9	2	3PLM	0.9009	0.9285	0.0576
MC10	2	3PLM	0.8985	0.4801	0.2104
MC11	2	3PLM	0.9799	1.1765	0.2381
MC12	2	3PLM	1.1006	-1.3998	0.1090
MC13	2	3PLM	0.9007	-1.6855	0.1976
MC14	2	3PLM	0.8600	1.6728	0.0417
MC15	2	3PLM	1.3216	-2.4808	0.2308
MC16	2	3PLM	0.6099	-1.2131	0.1369
MC17	2	3PLM	1.1016	-1.7968	0.1523
MC18	2	3PLM	0.7754	-1.7103	0.1522
MC19	2	3PLM	0.7330	-2.1310	0.2184
MC20	2	3PLM	1.1633	-0.9124	0.2428
MC21	2	3PLM	1.5057	-0.5905	0.0417
MC22	2	3PLM	0.9362	-0.5178	0.1682
MC23	2	3PLM	0.7434	1.4158	0.1683
MC24	2	3PLM	1.2617	-1.3794	0.2062
MC25	2	3PLM	1.0523	0.4533	0.1444
MC26	2	3PLM	0.9459	1.8983	0.2035
MC27	2	3PLM	1.9465	0.0296	0.2652
MC28	2	3PLM	1.2135	1.0579	0.0066
MC29	2	3PLM	0.9380	0.3072	0.2759
MC30	2	3PLM	0.8057	1.8552	0.0428
MC31	2	3PLM	1.0232	2.7537	0.1300
MC32	2	3PLM	0.8819	1.8000	0.2863
MC33	2	3PLM	0.6795	-0.3477	0.2258
MC34	2	3PLM	1.3259	-0.7747	0.1670
MC35	2	3PLM	0.8834	-0.1652	0.0543
MC36	2	3PLM	0.9771	1.9280	0.1251
MC37	2	3PLM	0.6141	0.2413	0.1592
MC38	2	3PLM	0.8588	-0.6472	0.2429
MC39	2	3PLM	1.2397	0.6192	0.0778
MC40	2	3PLM	1.2270	0.5034	0.2625

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.1460	-0.7217	0.1972
MC42	2	3PLM	1.0320	-1.1424	0.1972
MC43	2	3PLM	0.9178	-2.8722	0.2063
MC44	2	3PLM	0.5755	1.0393	0.0552
MC45	2	3PLM	1.0611	1.0851	0.1980
MC46	2	3PLM	1.2581	-0.1391	0.0290
MC47	2	3PLM	0.9653	0.0086	0.0863
MC48	2	3PLM	0.9184	-2.0288	0.1132
MC49	2	3PLM	0.9783	0.9686	0.0791
MC50	2	3PLM	0.9236	-0.3556	0.2108
MC51	2	3PLM	0.9605	1.0062	0.0700
MC52	2	3PLM	1.0917	1.2808	0.0000
MC53	2	3PLM	0.8163	0.8249	0.2075
MC54	2	3PLM	1.5170	0.4461	0.0992
FR1	3	GPCM	1.3403	0.9740	
FR2	3	GPCM	0.9169	-0.3557	
FR3	4	GPCM	1.5786	-0.1996	
FR4	4	GPCM	1.1920	-0.7630	
FR5	5	GPCM	0.7073	0.5467	
FR6	5	GPCM	0.9184	0.2538	

Table A21 IRT Item Parameters, NEAT, Design 1, Human Raters, Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0113	0.2804	0.2135
MC2	2	3PLM	0.8470	-0.4983	0.2055
MC3	2	3PLM	1.0260	0.2191	0.0602
MC4	2	3PLM	0.8686	1.0128	0.2319
MC5	2	3PLM	0.8511	0.3716	0.2854
MC6	2	3PLM	0.8711	0.5391	0.2046
MC7	2	3PLM	1.1054	0.6503	0.1716
MC8	2	3PLM	1.6024	1.0433	0.1929
MC9	2	3PLM	1.3567	0.1886	0.0431
MC10	2	3PLM	1.1392	-0.2847	0.2870
MC11	2	3PLM	1.0704	0.0831	0.0490
MC12	2	3PLM	1.5040	-0.1624	0.0263
MC13	2	3PLM	1.2833	1.1441	0.2646
MC14	2	3PLM	1.4017	1.0237	0.2991
MC15	2	3PLM	0.9282	-2.3662	0.2496
MC16	2	3PLM	1.1124	-0.7675	0.0474
MC17	2	3PLM	0.7514	-0.9673	0.1230
MC18	2	3PLM	0.8510	0.0647	0.1312
MC19	2	3PLM	1.2402	-0.3285	0.1791
MC20	2	3PLM	1.2010	0.2510	0.1012
MC21	2	3PLM	0.7130	-0.3392	0.1101
MC22	2	3PLM	1.0713	0.2934	0.2316
MC23	2	3PLM	1.1524	-0.9971	0.0861
MC24	2	3PLM	1.1155	-0.4834	0.1847
MC25	2	3PLM	1.4811	-0.7000	0.0832
MC26	2	3PLM	0.8887	0.1903	0.0799
MC27	2	3PLM	0.7777	0.0012	0.1754
MC28	2	3PLM	1.0224	0.2668	0.1388
MC29	2	3PLM	0.9535	-0.6696	0.1615
MC30	2	3PLM	1.0008	0.1816	0.2464
MC31	2	3PLM	0.9688	-2.0050	0.1708
MC32	2	3PLM	0.9980	-0.0285	0.0357
MC33	2	3PLM	0.9682	0.8396	0.2540
MC34	2	3PLM	1.3667	-0.2212	0.0608
MC35	2	3PLM	1.1592	0.5309	0.2367
MC36	2	3PLM	1.0551	0.0996	0.0651
MC37	2	3PLM	0.8913	0.4497	0.1444
MC38	2	3PLM	0.8745	-0.6699	0.0933
MC39	2	3PLM	1.5246	-0.9892	0.1768
MC40	2	3PLM	1.3038	-1.1696	0.3042

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.0563	-1.0318	0.1005
MC42	2	3PLM	1.5298	-0.5702	0.2668
MC43	2	3PLM	1.0682	-0.7432	0.1449
MC44	2	3PLM	1.1652	1.4432	0.2229
MC45	2	3PLM	1.3296	0.4742	0.2885
MC46	2	3PLM	0.9283	-0.5435	0.1843
MC47	2	3PLM	1.0334	0.5594	0.2416
MC48	2	3PLM	1.2004	-0.8579	0.0000
MC49	2	3PLM	0.9379	-1.0932	0.2459
MC50	2	3PLM	0.9600	-0.0850	0.2974
MC51	2	3PLM	0.8907	-0.8502	0.1311
MC52	2	3PLM	1.0120	0.1501	0.3301
MC53	2	3PLM	0.6840	1.9089	0.2981
MC54	2	3PLM	0.8778	-1.5945	0.3000
MC55	2	3PLM	0.9604	0.8732	0.1409
MC56	2	3PLM	1.4226	0.0434	0.1030
MC57	2	3PLM	0.9063	-0.4422	0.3233
FR1	3	GPCM	0.3519	-1.0164	
FR2	4	GPCM	0.6074	-0.7563	
FR3	5	GPCM	0.6080	0.0972	

Table A22 Equated IRT Item Parameters, NEAT, Design 1, Human Raters, Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.9911	2.4953	0.0772
MC2	2	3PLM	0.8901	-0.1844	0.1309
MC3	2	3PLM	1.2013	1.5466	0.2617
MC4	2	3PLM	1.1374	0.3566	0.1632
MC5	2	3PLM	1.3127	0.0635	0.0855
MC6	2	3PLM	1.0644	0.2742	0.2012
MC7	2	3PLM	1.0297	1.4092	0.0911
MC8	2	3PLM	0.8499	-1.3952	0.1750
MC9	2	3PLM	0.9255	-1.0502	0.1162
MC10	2	3PLM	0.9151	0.3130	0.1051
MC11	2	3PLM	1.7355	1.4045	0.0554
MC12	2	3PLM	1.1640	0.4000	0.0332
MC13	2	3PLM	1.1130	0.6021	0.0653
MC14	2	3PLM	1.0988	-0.3037	0.1288
MC15	2	3PLM	0.8716	-0.8926	0.1740
MC16	2	3PLM	0.8651	0.4707	0.1084
MC17	2	3PLM	0.9119	0.4046	0.0807
MC18	2	3PLM	1.0209	-1.2632	0.1584
MC19	2	3PLM	0.2779	-0.0760	0.0939
MC20	2	3PLM	0.8552	1.8159	0.3095
MC21	2	3PLM	1.1165	0.7179	0.0979
MC22	2	3PLM	0.9538	0.4891	0.1966
MC23	2	3PLM	1.2152	-1.3005	0.2093
MC24	2	3PLM	1.3390	-0.3877	0.0634
MC25	2	3PLM	1.1754	1.8481	0.2998
MC26	2	3PLM	0.8820	0.0209	0.0553
MC27	2	3PLM	0.8920	1.8074	0.0442
MC28	2	3PLM	1.0519	-0.4336	0.2818
MC29	2	3PLM	1.0875	-0.5003	0.2338
MC30	2	3PLM	0.9212	-0.7465	0.1257
MC31	2	3PLM	0.9082	-2.7169	0.2320
MC32	2	3PLM	1.0726	-1.3865	0.1998
MC33	2	3PLM	1.3606	0.2279	0.0367
MC34	2	3PLM	1.1854	-1.2375	0.1618
MC35	2	3PLM	0.9658	0.7840	0.1233
MC36	2	3PLM	1.1837	-0.8861	0.0855
MC37	2	3PLM	1.0009	0.3590	0.2239
MC38	2	3PLM	1.0728	0.2764	0.0690
MC39	2	3PLM	1.1150	0.7226	0.1645
MC40	2	3PLM	1.3337	0.2408	0.0440

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.1682	0.1479	0.0724
MC42	2	3PLM	1.3949	1.1615	0.2474
MC43	2	3PLM	0.8891	-2.5100	0.1879
MC44	2	3PLM	1.2170	-0.3027	0.1623
MC45	2	3PLM	0.6891	-0.2610	0.1197
MC46	2	3PLM	1.5285	-0.6386	0.0688
MC47	2	3PLM	0.7358	-0.0871	0.1223
MC48	2	3PLM	1.0566	-1.8404	0.1969
MC49	2	3PLM	0.9863	0.9150	0.2569
MC50	2	3PLM	0.9772	0.5196	0.1726
MC51	2	3PLM	1.5381	-0.8531	0.2294
MC52	2	3PLM	1.2119	-0.5533	0.2211
MC53	2	3PLM	1.0079	0.4681	0.2561
MC54	2	3PLM	1.1432	0.6134	0.2560
MC55	2	3PLM	0.9045	-0.7921	0.1480
MC56	2	3PLM	0.7753	1.8892	0.3068
MC57	2	3PLM	0.9146	-0.4472	0.2958
FR1	3	GPCM	0.4174	-1.0730	
FR2	4	GPCM	0.5381	-0.9102	
FR3	5	GPCM	0.6571	0.0076	

Table A23 IRT Item Parameters, NEAT, Design 2, Human Raters, Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.6724	0.3459	0.1866
MC2	2	3PLM	0.8487	-1.6661	0.1978
MC3	2	3PLM	0.8661	-0.6716	0.2703
MC4	2	3PLM	0.8181	-0.6712	0.1081
MC5	2	3PLM	1.2093	0.6770	0.0918
MC6	2	3PLM	0.9908	-1.2569	0.1883
MC7	2	3PLM	1.0474	0.6773	0.2459
MC8	2	3PLM	0.8406	-0.4866	0.3216
MC9	2	3PLM	1.3254	0.5268	0.2761
MC10	2	3PLM	0.7885	2.3448	0.1662
MC11	2	3PLM	1.1430	-0.7356	0.2010
MC12	2	3PLM	1.1373	1.1947	0.1813
MC13	2	3PLM	1.0638	-1.1339	0.1764
MC14	2	3PLM	0.7711	-1.0205	0.0998
MC15	2	3PLM	1.5331	0.2450	0.0443
MC16	2	3PLM	0.9051	-2.0584	0.0000
MC17	2	3PLM	0.9467	-2.8595	0.2283
MC18	2	3PLM	0.9200	-1.0594	0.0851
MC19	2	3PLM	0.6206	1.0371	0.0778
MC20	2	3PLM	0.8189	-0.4845	0.0961
MC21	2	3PLM	0.7795	-1.6514	0.0000
MC22	2	3PLM	0.8210	0.2377	0.2345
MC23	2	3PLM	1.0680	1.0710	0.2013
MC24	2	3PLM	1.0521	-0.2559	0.0578
MC25	2	3PLM	1.0493	-1.5221	0.0000
MC26	2	3PLM	0.7777	0.6813	0.0888
MC27	2	3PLM	1.1062	0.1838	0.2126
MC28	2	3PLM	0.8439	-1.5644	0.1398
MC29	2	3PLM	1.2244	-0.1318	0.0290
MC30	2	3PLM	1.2126	0.9427	0.0140
MC31	2	3PLM	0.9588	0.0369	0.1071
MC32	2	3PLM	0.8342	-0.4589	0.1750
MC33	2	3PLM	0.9248	-1.9835	0.1561
MC34	2	3PLM	1.1030	0.7319	0.2328
MC35	2	3PLM	1.0535	-1.7877	0.1803
MC36	2	3PLM	1.1188	0.8957	0.1895
MC37	2	3PLM	1.0405	1.0178	0.0972
MC38	2	3PLM	1.1544	-0.2510	0.0439
MC39	2	3PLM	0.9509	-0.2837	0.2068
MC40	2	3PLM	0.8506	-0.0342	0.1476

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.4351	1.3439	0.0239
MC42	2	3PLM	1.1906	-0.7290	0.0473
MC43	2	3PLM	1.0430	0.9762	0.0851
MC44	2	3PLM	0.7237	-0.7659	0.0000
MC45	2	3PLM	1.1942	1.2449	0.0098
MC46	2	3PLM	0.8303	-0.0720	0.3057
MC47	2	3PLM	0.8863	0.8631	0.2096
MC48	2	3PLM	1.1808	-1.2315	0.1439
MC49	2	3PLM	1.4050	0.5822	0.0338
MC50	2	3PLM	1.0667	-0.2183	0.2527
MC51	2	3PLM	1.0335	-0.4992	0.2180
MC52	2	3PLM	0.7739	-1.6641	0.1662
MC53	2	3PLM	1.3682	0.4135	0.0953
MC54	2	3PLM	0.7595	1.4131	0.1575
FR1	3	GPCM	0.6339	1.0509	
FR2	3	GPCM	0.5863	-0.3548	
FR3	4	GPCM	0.7096	-0.1762	
FR4	4	GPCM	0.6673	-0.8014	
FR5	5	GPCM	0.4680	0.5770	
FR6	5	GPCM	0.6273	0.2335	

Table A24 Equated IRT Item Parameters, NEAT, Design 2, Human Raters, Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.2344	-1.4367	0.1687
MC2	2	3PLM	0.9239	1.6831	0.2916
MC3	2	3PLM	0.7195	-1.1553	0.2019
MC4	2	3PLM	0.7962	-0.5155	0.1917
MC5	2	3PLM	1.2292	-1.0352	0.1129
MC6	2	3PLM	1.3468	0.7662	0.2619
MC7	2	3PLM	1.0587	-0.3609	0.0373
MC8	2	3PLM	0.8918	-0.9780	0.2215
MC9	2	3PLM	0.8928	0.9275	0.0569
MC10	2	3PLM	0.8925	0.4791	0.2108
MC11	2	3PLM	0.9839	1.1691	0.2376
MC12	2	3PLM	1.1198	-1.3956	0.1089
MC13	2	3PLM	0.9051	-1.7028	0.1829
MC14	2	3PLM	0.8815	1.6580	0.0435
MC15	2	3PLM	1.4042	-2.4056	0.2320
MC16	2	3PLM	0.6068	-1.2378	0.1281
MC17	2	3PLM	1.1186	-1.7929	0.1446
MC18	2	3PLM	0.7818	-1.7044	0.1533
MC19	2	3PLM	0.7421	-2.1335	0.2053
MC20	2	3PLM	1.1570	-0.9307	0.2376
MC21	2	3PLM	1.5004	-0.6037	0.0405
MC22	2	3PLM	0.9239	-0.5390	0.1622
MC23	2	3PLM	0.7624	1.4046	0.1706
MC24	2	3PLM	1.2606	-1.3982	0.1952
MC25	2	3PLM	1.0561	0.4612	0.1490
MC26	2	3PLM	0.9963	1.8766	0.2068
MC27	2	3PLM	1.8929	0.0166	0.2620
MC28	2	3PLM	1.2205	1.0576	0.0072
MC29	2	3PLM	0.9290	0.2991	0.2743
MC30	2	3PLM	0.8246	1.8418	0.0449
MC31	2	3PLM	1.0643	2.6857	0.1300
MC32	2	3PLM	0.9188	1.7871	0.2895
MC33	2	3PLM	0.6710	-0.3823	0.2153
MC34	2	3PLM	1.3267	-0.7875	0.1650
MC35	2	3PLM	0.8735	-0.1785	0.0512
MC36	2	3PLM	0.9908	1.9153	0.1256
MC37	2	3PLM	0.6070	0.2282	0.1561
MC38	2	3PLM	0.8474	-0.6736	0.2355
MC39	2	3PLM	1.2325	0.6224	0.0793
MC40	2	3PLM	1.2255	0.5056	0.2639

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.1469	-0.7260	0.2003
MC42	2	3PLM	1.0279	-1.1650	0.1875
MC43	2	3PLM	0.9297	-2.8447	0.2015
MC44	2	3PLM	0.5753	1.0323	0.0546
MC45	2	3PLM	1.0688	1.0870	0.1995
MC46	2	3PLM	1.2328	-0.1498	0.0277
MC47	2	3PLM	0.9512	-0.0047	0.0828
MC48	2	3PLM	0.9329	-2.0069	0.1186
MC49	2	3PLM	0.9725	0.9694	0.0792
MC50	2	3PLM	0.9123	-0.3804	0.2031
MC51	2	3PLM	0.9686	1.0094	0.0725
MC52	2	3PLM	1.0815	1.2756	0.0000
MC53	2	3PLM	0.8275	0.8343	0.2123
MC54	2	3PLM	1.4879	0.4432	0.0979
FR1	3	GPCM	0.6632	1.0123	
FR2	3	GPCM	0.6051	-0.3600	
FR3	4	GPCM	0.7171	-0.1358	
FR4	4	GPCM	0.6668	-0.7890	
FR5	5	GPCM	0.4667	0.5636	
FR6	5	GPCM	0.6070	0.2338	

Table A25 IRT Item Parameters, NEAT, Design 1, Automated Raters (Constant Noise), Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0113	0.2804	0.2135
MC2	2	3PLM	0.8470	-0.4983	0.2055
MC3	2	3PLM	1.0260	0.2191	0.0602
MC4	2	3PLM	0.8686	1.0128	0.2319
MC5	2	3PLM	0.8511	0.3716	0.2854
MC6	2	3PLM	0.8711	0.5391	0.2046
MC7	2	3PLM	1.1054	0.6503	0.1716
MC8	2	3PLM	1.6024	1.0433	0.1929
MC9	2	3PLM	1.3567	0.1886	0.0431
MC10	2	3PLM	1.1392	-0.2847	0.2870
MC11	2	3PLM	1.0704	0.0831	0.0490
MC12	2	3PLM	1.5040	-0.1624	0.0263
MC13	2	3PLM	1.2833	1.1441	0.2646
MC14	2	3PLM	1.4017	1.0237	0.2991
MC15	2	3PLM	0.9282	-2.3662	0.2496
MC16	2	3PLM	1.1124	-0.7675	0.0474
MC17	2	3PLM	0.7514	-0.9673	0.1230
MC18	2	3PLM	0.8510	0.0647	0.1312
MC19	2	3PLM	1.2402	-0.3285	0.1791
MC20	2	3PLM	1.2010	0.2510	0.1012
MC21	2	3PLM	0.7130	-0.3392	0.1101
MC22	2	3PLM	1.0713	0.2934	0.2316
MC23	2	3PLM	1.1524	-0.9971	0.0861
MC24	2	3PLM	1.1155	-0.4834	0.1847
MC25	2	3PLM	1.4811	-0.7000	0.0832
MC26	2	3PLM	0.8887	0.1903	0.0799
MC27	2	3PLM	0.7777	0.0012	0.1754
MC28	2	3PLM	1.0224	0.2668	0.1388
MC29	2	3PLM	0.9535	-0.6696	0.1615
MC30	2	3PLM	1.0008	0.1816	0.2464
MC31	2	3PLM	0.9688	-2.0050	0.1708
MC32	2	3PLM	0.9980	-0.0285	0.0357
MC33	2	3PLM	0.9682	0.8396	0.2540
MC34	2	3PLM	1.3667	-0.2212	0.0608
MC35	2	3PLM	1.1592	0.5309	0.2367
MC36	2	3PLM	1.0551	0.0996	0.0651
MC37	2	3PLM	0.8913	0.4497	0.1444
MC38	2	3PLM	0.8745	-0.6699	0.0933
MC39	2	3PLM	1.5246	-0.9892	0.1768
MC40	2	3PLM	1.3038	-1.1696	0.3042

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.0563	-1.0318	0.1005
MC42	2	3PLM	1.5298	-0.5702	0.2668
MC43	2	3PLM	1.0682	-0.7432	0.1449
MC44	2	3PLM	1.1652	1.4432	0.2229
MC45	2	3PLM	1.3296	0.4742	0.2885
MC46	2	3PLM	0.9283	-0.5435	0.1843
MC47	2	3PLM	1.0334	0.5594	0.2416
MC48	2	3PLM	1.2004	-0.8579	0.0000
MC49	2	3PLM	0.9379	-1.0932	0.2459
MC50	2	3PLM	0.9600	-0.0850	0.2974
MC51	2	3PLM	0.8907	-0.8502	0.1311
MC52	2	3PLM	1.0120	0.1501	0.3301
MC53	2	3PLM	0.6840	1.9089	0.2981
MC54	2	3PLM	0.8778	-1.5945	0.3000
MC55	2	3PLM	0.9604	0.8732	0.1409
MC56	2	3PLM	1.4226	0.0434	0.1030
MC57	2	3PLM	0.9063	-0.4422	0.3233
FR1	3	GPCM	0.3519	-1.0164	
FR2	4	GPCM	0.6074	-0.7563	
FR3	5	GPCM	0.6080	0.0972	

Table A26 Equated IRT Item Parameters, NEAT, Design 1, Automated Raters (Constant Noise), Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.9251	2.6595	0.0769
MC2	2	3PLM	0.8445	-0.1809	0.1370
MC3	2	3PLM	1.1434	1.6443	0.2627
MC4	2	3PLM	1.0812	0.3838	0.1657
MC5	2	3PLM	1.2316	0.0633	0.0833
MC6	2	3PLM	0.9941	0.2853	0.1985
MC7	2	3PLM	0.9805	1.4980	0.0924
MC8	2	3PLM	0.7945	-1.5047	0.1619
MC9	2	3PLM	0.8713	-1.1124	0.1170
MC10	2	3PLM	0.8616	0.3253	0.1024
MC11	2	3PLM	1.6378	1.4935	0.0557
MC12	2	3PLM	1.0915	0.4263	0.0335
MC13	2	3PLM	1.0634	0.6459	0.0689
MC14	2	3PLM	1.0360	-0.3239	0.1273
MC15	2	3PLM	0.8217	-0.9510	0.1708
MC16	2	3PLM	0.8078	0.5006	0.1078
MC17	2	3PLM	0.8593	0.4286	0.0799
MC18	2	3PLM	0.9620	-1.3347	0.1625
MC19	2	3PLM	0.2724	-0.0830	0.1137
MC20	2	3PLM	0.8118	1.9337	0.3105
MC21	2	3PLM	1.0515	0.7636	0.0982
MC22	2	3PLM	0.8937	0.5179	0.1958
MC23	2	3PLM	1.1600	-1.3649	0.2140
MC24	2	3PLM	1.2604	-0.4143	0.0607
MC25	2	3PLM	1.0969	1.9713	0.2997
MC26	2	3PLM	0.8238	0.0171	0.0524
MC27	2	3PLM	0.8311	1.9273	0.0436
MC28	2	3PLM	0.9978	-0.4537	0.2837
MC29	2	3PLM	1.0222	-0.5307	0.2339
MC30	2	3PLM	0.8613	-0.8114	0.1152
MC31	2	3PLM	0.8608	-2.8767	0.2254
MC32	2	3PLM	1.0204	-1.4560	0.2050
MC33	2	3PLM	1.2789	0.2413	0.0358
MC34	2	3PLM	1.1079	-1.3214	0.1580
MC35	2	3PLM	0.9126	0.8361	0.1244
MC36	2	3PLM	1.1161	-0.9355	0.0884
MC37	2	3PLM	0.9409	0.3846	0.2245
MC38	2	3PLM	1.0072	0.2925	0.0686
MC39	2	3PLM	1.0415	0.7675	0.1638

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	1.2703	0.2595	0.0463
MC41	2	3PLM	1.1095	0.1604	0.0743
MC42	2	3PLM	1.3144	1.2347	0.2477
MC43	2	3PLM	0.8397	-2.6677	0.1792
MC44	2	3PLM	1.1368	-0.3305	0.1570
MC45	2	3PLM	0.6434	-0.2953	0.1122
MC46	2	3PLM	1.4324	-0.6824	0.0650
MC47	2	3PLM	0.6926	-0.1014	0.1180
MC48	2	3PLM	0.9997	-1.9486	0.1943
MC49	2	3PLM	0.9291	0.9719	0.2568
MC50	2	3PLM	0.9260	0.5548	0.1739
MC51	2	3PLM	1.4595	-0.8995	0.2312
MC52	2	3PLM	1.1442	-0.5860	0.2214
MC53	2	3PLM	0.9545	0.5043	0.2585
MC54	2	3PLM	1.0665	0.6440	0.2534
MC55	2	3PLM	0.8547	-0.8342	0.1504
MC56	2	3PLM	0.7202	2.0107	0.3054
MC57	2	3PLM	0.8544	-0.4900	0.2898
FR1	3	GPCM	0.5337	-1.1669	
FR2	4	GPCM	0.9176	-0.8471	
FR3	5	GPCM	0.8213	0.0833	

Table A27 IRT Item Parameters, NEAT, Design 2, Automated Raters (Constant Noise), Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.6811	0.3612	0.1891
MC2	2	3PLM	0.8377	-1.6632	0.2043
MC3	2	3PLM	0.8656	-0.6611	0.2698
MC4	2	3PLM	0.8215	-0.6470	0.1157
MC5	2	3PLM	1.2399	0.6860	0.0949
MC6	2	3PLM	0.9831	-1.2562	0.1857
MC7	2	3PLM	1.0871	0.6876	0.2503
MC8	2	3PLM	0.8197	-0.5128	0.3090
MC9	2	3PLM	1.3682	0.5384	0.2783
MC10	2	3PLM	0.8135	2.3092	0.1672
MC11	2	3PLM	1.1660	-0.6937	0.2179
MC12	2	3PLM	1.1663	1.1872	0.1815
MC13	2	3PLM	1.0861	-1.0754	0.2077
MC14	2	3PLM	0.7769	-0.9910	0.1105
MC15	2	3PLM	1.5468	0.2552	0.0440
MC16	2	3PLM	0.9176	-1.9173	0.0000
MC17	2	3PLM	0.9446	-2.8697	0.2193
MC18	2	3PLM	0.9177	-1.0570	0.0815
MC19	2	3PLM	0.6361	1.0478	0.0829
MC20	2	3PLM	0.8254	-0.4625	0.1020
MC21	2	3PLM	0.7718	-1.6790	0.0000
MC22	2	3PLM	0.8475	0.2685	0.2438
MC23	2	3PLM	1.0739	1.0649	0.1995
MC24	2	3PLM	1.0604	-0.2453	0.0555
MC25	2	3PLM	1.0889	-1.3913	0.0968
MC26	2	3PLM	0.7853	0.6885	0.0898
MC27	2	3PLM	1.1520	0.2146	0.2230
MC28	2	3PLM	0.8357	-1.5734	0.1354
MC29	2	3PLM	1.2325	-0.1165	0.0305
MC30	2	3PLM	1.2231	0.9448	0.0139
MC31	2	3PLM	0.9737	0.0537	0.1107
MC32	2	3PLM	0.8340	-0.4558	0.1714
MC33	2	3PLM	0.9181	-2.0028	0.1371
MC34	2	3PLM	1.1230	0.7344	0.2326
MC35	2	3PLM	1.0344	-1.8032	0.1743
MC36	2	3PLM	1.1518	0.8985	0.1917
MC37	2	3PLM	1.0591	1.0196	0.0984
MC38	2	3PLM	1.1541	-0.2439	0.0408
MC39	2	3PLM	0.9521	-0.2684	0.2095
MC40	2	3PLM	0.8533	-0.0262	0.1461

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.4625	1.3360	0.0237
MC42	2	3PLM	1.1961	-0.7168	0.0477
MC43	2	3PLM	1.0671	0.9796	0.0874
MC44	2	3PLM	0.7650	-0.6346	0.0605
MC45	2	3PLM	1.2154	1.2403	0.0103
MC46	2	3PLM	0.8481	-0.0415	0.3129
MC47	2	3PLM	0.9044	0.8603	0.2094
MC48	2	3PLM	1.1962	-1.1893	0.1688
MC49	2	3PLM	1.4222	0.5928	0.0356
MC50	2	3PLM	1.0782	-0.2044	0.2534
MC51	2	3PLM	1.0603	-0.4513	0.2352
MC52	2	3PLM	0.7679	-1.6666	0.1626
MC53	2	3PLM	1.3832	0.4246	0.0963
MC54	2	3PLM	0.7933	1.4001	0.1611
FR1	3	GPCM	0.6987	1.1852	
FR2	3	GPCM	0.5754	-0.1764	
FR3	4	GPCM	0.7879	-0.0002	
FR4	4	GPCM	0.6943	-0.6524	
FR5	5	GPCM	0.4346	0.7731	
FR6	5	GPCM	0.5829	0.3850	

Table A28 Equated IRT Item Parameters, NEAT, Design 2, Automated Raters (Constant Noise), Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0841	-1.5569	0.2017
MC2	2	3PLM	0.7968	1.9580	0.2899
MC3	2	3PLM	0.6487	-1.1809	0.2372
MC4	2	3PLM	0.7102	-0.5121	0.2032
MC5	2	3PLM	1.0891	-1.1164	0.1228
MC6	2	3PLM	1.1916	0.9124	0.2623
MC7	2	3PLM	0.9291	-0.3663	0.0362
MC8	2	3PLM	0.7881	-1.0503	0.2293
MC9	2	3PLM	0.8024	1.0972	0.0598
MC10	2	3PLM	0.7892	0.5854	0.2102
MC11	2	3PLM	0.8673	1.3739	0.2383
MC12	2	3PLM	0.9611	-1.5666	0.1063
MC13	2	3PLM	0.7889	-1.8891	0.1952
MC14	2	3PLM	0.7677	1.9322	0.0427
MC15	2	3PLM	1.1518	-2.8057	0.2245
MC16	2	3PLM	0.5322	-1.3510	0.1371
MC17	2	3PLM	0.9659	-2.0167	0.1485
MC18	2	3PLM	0.6779	-1.9198	0.1515
MC19	2	3PLM	0.6409	-2.4102	0.2116
MC20	2	3PLM	1.0150	-1.0071	0.2431
MC21	2	3PLM	1.3167	-0.6361	0.0425
MC22	2	3PLM	0.8123	-0.5666	0.1630
MC23	2	3PLM	0.6687	1.6420	0.1705
MC24	2	3PLM	1.1058	-1.5443	0.2005
MC25	2	3PLM	0.9258	0.5595	0.1456
MC26	2	3PLM	0.8505	2.1799	0.2042
MC27	2	3PLM	1.6937	0.0751	0.2657
MC28	2	3PLM	1.0725	1.2420	0.0068
MC29	2	3PLM	0.8289	0.3945	0.2778
MC30	2	3PLM	0.7284	2.1314	0.0449
MC31	2	3PLM	0.8799	3.1272	0.1297
MC32	2	3PLM	0.8009	2.0695	0.2883
MC33	2	3PLM	0.5915	-0.3788	0.2196
MC34	2	3PLM	1.1634	-0.8401	0.1717
MC35	2	3PLM	0.7724	-0.1537	0.0530
MC36	2	3PLM	0.8711	2.2171	0.1255
MC37	2	3PLM	0.5402	0.3142	0.1600
MC38	2	3PLM	0.7524	-0.6964	0.2449
MC39	2	3PLM	1.0919	0.7490	0.0792

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	1.0917	0.6197	0.2651
MC41	2	3PLM	1.0086	-0.7765	0.2028
MC42	2	3PLM	0.8973	-1.2746	0.1955
MC43	2	3PLM	0.8087	-3.2334	0.2021
MC44	2	3PLM	0.5117	1.2199	0.0572
MC45	2	3PLM	0.9473	1.2744	0.1994
MC46	2	3PLM	1.0984	-0.1220	0.0285
MC47	2	3PLM	0.8484	0.0510	0.0872
MC48	2	3PLM	0.8058	-2.2727	0.1148
MC49	2	3PLM	0.8567	1.1394	0.0782
MC50	2	3PLM	0.8123	-0.3665	0.2115
MC51	2	3PLM	0.8573	1.1870	0.0724
MC52	2	3PLM	0.9647	1.4933	0.0000
MC53	2	3PLM	0.7272	0.9836	0.2108
MC54	2	3PLM	1.3346	0.5531	0.1010
FR1	3	GPCM	1.0530	1.1816	
FR2	3	GPCM	0.7246	-0.3348	
FR3	4	GPCM	1.1636	-0.1351	
FR4	4	GPCM	0.8851	-0.8215	
FR5	5	GPCM	0.5330	0.7706	
FR6	5	GPCM	0.6755	0.4022	

Table A29 IRT Item Parameters, NEAT, Design 1, Automated Raters (Reduced Noise), Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.0113	0.2804	0.2135
MC2	2	3PLM	0.8470	-0.4983	0.2055
MC3	2	3PLM	1.0260	0.2191	0.0602
MC4	2	3PLM	0.8686	1.0128	0.2319
MC5	2	3PLM	0.8511	0.3716	0.2854
MC6	2	3PLM	0.8711	0.5391	0.2046
MC7	2	3PLM	1.1054	0.6503	0.1716
MC8	2	3PLM	1.6024	1.0433	0.1929
MC9	2	3PLM	1.3567	0.1886	0.0431
MC10	2	3PLM	1.1392	-0.2847	0.2870
MC11	2	3PLM	1.0704	0.0831	0.0490
MC12	2	3PLM	1.5040	-0.1624	0.0263
MC13	2	3PLM	1.2833	1.1441	0.2646
MC14	2	3PLM	1.4017	1.0237	0.2991
MC15	2	3PLM	0.9282	-2.3662	0.2496
MC16	2	3PLM	1.1124	-0.7675	0.0474
MC17	2	3PLM	0.7514	-0.9673	0.1230
MC18	2	3PLM	0.8510	0.0647	0.1312
MC19	2	3PLM	1.2402	-0.3285	0.1791
MC20	2	3PLM	1.2010	0.2510	0.1012
MC21	2	3PLM	0.7130	-0.3392	0.1101
MC22	2	3PLM	1.0713	0.2934	0.2316
MC23	2	3PLM	1.1524	-0.9971	0.0861
MC24	2	3PLM	1.1155	-0.4834	0.1847
MC25	2	3PLM	1.4811	-0.7000	0.0832
MC26	2	3PLM	0.8887	0.1903	0.0799
MC27	2	3PLM	0.7777	0.0012	0.1754
MC28	2	3PLM	1.0224	0.2668	0.1388
MC29	2	3PLM	0.9535	-0.6696	0.1615
MC30	2	3PLM	1.0008	0.1816	0.2464
MC31	2	3PLM	0.9688	-2.0050	0.1708
MC32	2	3PLM	0.9980	-0.0285	0.0357
MC33	2	3PLM	0.9682	0.8396	0.2540
MC34	2	3PLM	1.3667	-0.2212	0.0608
MC35	2	3PLM	1.1592	0.5309	0.2367
MC36	2	3PLM	1.0551	0.0996	0.0651
MC37	2	3PLM	0.8913	0.4497	0.1444
MC38	2	3PLM	0.8745	-0.6699	0.0933
MC39	2	3PLM	1.5246	-0.9892	0.1768
MC40	2	3PLM	1.3038	-1.1696	0.3042

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.0563	-1.0318	0.1005
MC42	2	3PLM	1.5298	-0.5702	0.2668
MC43	2	3PLM	1.0682	-0.7432	0.1449
MC44	2	3PLM	1.1652	1.4432	0.2229
MC45	2	3PLM	1.3296	0.4742	0.2885
MC46	2	3PLM	0.9283	-0.5435	0.1843
MC47	2	3PLM	1.0334	0.5594	0.2416
MC48	2	3PLM	1.2004	-0.8579	0.0000
MC49	2	3PLM	0.9379	-1.0932	0.2459
MC50	2	3PLM	0.9600	-0.0850	0.2974
MC51	2	3PLM	0.8907	-0.8502	0.1311
MC52	2	3PLM	1.0120	0.1501	0.3301
MC53	2	3PLM	0.6840	1.9089	0.2981
MC54	2	3PLM	0.8778	-1.5945	0.3000
MC55	2	3PLM	0.9604	0.8732	0.1409
MC56	2	3PLM	1.4226	0.0434	0.1030
MC57	2	3PLM	0.9063	-0.4422	0.3233
FR1	3	GPCM	0.3519	-1.0164	
FR2	4	GPCM	0.6074	-0.7563	
FR3	5	GPCM	0.6080	0.0972	

Table A30 Equated IRT Item Parameters, NEAT, Design 1, Automated Raters (Reduced Noise), Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.9240	2.6552	0.0768
MC2	2	3PLM	0.8355	-0.1933	0.1323
MC3	2	3PLM	1.1810	1.6329	0.2640
MC4	2	3PLM	1.0870	0.3871	0.1676
MC5	2	3PLM	1.2313	0.0630	0.0834
MC6	2	3PLM	1.0013	0.2886	0.2003
MC7	2	3PLM	0.9878	1.4933	0.0927
MC8	2	3PLM	0.7983	-1.4987	0.1634
MC9	2	3PLM	0.8740	-1.1073	0.1195
MC10	2	3PLM	0.8517	0.3221	0.1006
MC11	2	3PLM	1.6265	1.4944	0.0554
MC12	2	3PLM	1.0968	0.4260	0.0341
MC13	2	3PLM	1.0560	0.6441	0.0677
MC14	2	3PLM	1.0329	-0.3330	0.1232
MC15	2	3PLM	0.8208	-0.9576	0.1673
MC16	2	3PLM	0.8134	0.5004	0.1085
MC17	2	3PLM	0.8528	0.4243	0.0783
MC18	2	3PLM	0.9689	-1.3233	0.1675
MC19	2	3PLM	0.2556	-0.0770	0.0883
MC20	2	3PLM	0.7967	1.9336	0.3087
MC21	2	3PLM	1.0480	0.7602	0.0969
MC22	2	3PLM	0.8926	0.5200	0.1964
MC23	2	3PLM	1.1421	-1.3965	0.1955
MC24	2	3PLM	1.2582	-0.4160	0.0600
MC25	2	3PLM	1.0768	1.9816	0.2993
MC26	2	3PLM	0.8225	0.0154	0.0520
MC27	2	3PLM	0.8380	1.9197	0.0439
MC28	2	3PLM	0.9773	-0.4813	0.2731
MC29	2	3PLM	1.0137	-0.5425	0.2288
MC30	2	3PLM	0.8585	-0.8158	0.1138
MC31	2	3PLM	0.8510	-2.8984	0.2242
MC32	2	3PLM	1.0181	-1.4686	0.1965
MC33	2	3PLM	1.2788	0.2397	0.0354
MC34	2	3PLM	1.1120	-1.3154	0.1613
MC35	2	3PLM	0.9164	0.8352	0.1248
MC36	2	3PLM	1.1141	-0.9428	0.0836
MC37	2	3PLM	0.9539	0.3935	0.2285
MC38	2	3PLM	1.0125	0.2946	0.0698
MC39	2	3PLM	1.0550	0.7693	0.1656

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	1.2693	0.2577	0.0457
MC41	2	3PLM	1.1070	0.1568	0.0726
MC42	2	3PLM	1.3046	1.2356	0.2475
MC43	2	3PLM	0.8445	-2.6538	0.1835
MC44	2	3PLM	1.1453	-0.3238	0.1605
MC45	2	3PLM	0.6433	-0.3022	0.1093
MC46	2	3PLM	1.4393	-0.6805	0.0663
MC47	2	3PLM	0.6939	-0.0962	0.1204
MC48	2	3PLM	0.9970	-1.9608	0.1861
MC49	2	3PLM	0.9334	0.9696	0.2570
MC50	2	3PLM	0.9297	0.5570	0.1752
MC51	2	3PLM	1.4512	-0.9057	0.2282
MC52	2	3PLM	1.1303	-0.5934	0.2193
MC53	2	3PLM	0.9466	0.5017	0.2574
MC54	2	3PLM	1.0744	0.6481	0.2554
MC55	2	3PLM	0.8541	-0.8313	0.1526
MC56	2	3PLM	0.7187	2.0080	0.3052
MC57	2	3PLM	0.8588	-0.4784	0.2946
FR1	3	GPCM	0.5207	-1.1852	
FR2	4	GPCM	0.8986	-0.8451	
FR3	5	GPCM	0.8170	0.0859	

Table A31IRT Item Parameters, NEAT, Design 2, Automated Raters (Reduced Noise), Form 1

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	0.6811	0.3612	0.1891
MC2	2	3PLM	0.8377	-1.6632	0.2043
MC3	2	3PLM	0.8656	-0.6611	0.2698
MC4	2	3PLM	0.8215	-0.6470	0.1157
MC5	2	3PLM	1.2399	0.6860	0.0949
MC6	2	3PLM	0.9831	-1.2562	0.1857
MC7	2	3PLM	1.0871	0.6876	0.2503
MC8	2	3PLM	0.8197	-0.5128	0.3090
MC9	2	3PLM	1.3682	0.5384	0.2783
MC10	2	3PLM	0.8135	2.3092	0.1672
MC11	2	3PLM	1.1660	-0.6937	0.2179
MC12	2	3PLM	1.1663	1.1872	0.1815
MC13	2	3PLM	1.0861	-1.0754	0.2077
MC14	2	3PLM	0.7769	-0.9910	0.1105
MC15	2	3PLM	1.5468	0.2552	0.0440
MC16	2	3PLM	0.9176	-1.9173	0.0000
MC17	2	3PLM	0.9446	-2.8697	0.2193
MC18	2	3PLM	0.9177	-1.0570	0.0815
MC19	2	3PLM	0.6361	1.0478	0.0829
MC20	2	3PLM	0.8254	-0.4625	0.1020
MC21	2	3PLM	0.7718	-1.6790	0.0000
MC22	2	3PLM	0.8475	0.2685	0.2438
MC23	2	3PLM	1.0739	1.0649	0.1995
MC24	2	3PLM	1.0604	-0.2453	0.0555
MC25	2	3PLM	1.0889	-1.3913	0.0968
MC26	2	3PLM	0.7853	0.6885	0.0898
MC27	2	3PLM	1.1520	0.2146	0.2230
MC28	2	3PLM	0.8357	-1.5734	0.1354
MC29	2	3PLM	1.2325	-0.1165	0.0305
MC30	2	3PLM	1.2231	0.9448	0.0139
MC31	2	3PLM	0.9737	0.0537	0.1107
MC32	2	3PLM	0.8340	-0.4558	0.1714
MC33	2	3PLM	0.9181	-2.0028	0.1371
MC34	2	3PLM	1.1230	0.7344	0.2326
MC35	2	3PLM	1.0344	-1.8032	0.1743
MC36	2	3PLM	1.1518	0.8985	0.1917
MC37	2	3PLM	1.0591	1.0196	0.0984
MC38	2	3PLM	1.1541	-0.2439	0.0408
MC39	2	3PLM	0.9521	-0.2684	0.2095
MC40	2	3PLM	0.8533	-0.0262	0.1461

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC41	2	3PLM	1.4625	1.3360	0.0237
MC42	2	3PLM	1.1961	-0.7168	0.0477
MC43	2	3PLM	1.0671	0.9796	0.0874
MC44	2	3PLM	0.7650	-0.6346	0.0605
MC45	2	3PLM	1.2154	1.2403	0.0103
MC46	2	3PLM	0.8481	-0.0415	0.3129
MC47	2	3PLM	0.9044	0.8603	0.2094
MC48	2	3PLM	1.1962	-1.1893	0.1688
MC49	2	3PLM	1.4222	0.5928	0.0356
MC50	2	3PLM	1.0782	-0.2044	0.2534
MC51	2	3PLM	1.0603	-0.4513	0.2352
MC52	2	3PLM	0.7679	-1.6666	0.1626
MC53	2	3PLM	1.3832	0.4246	0.0963
MC54	2	3PLM	0.7933	1.4001	0.1611
FR1	3	GPCM	0.6987	1.1852	
FR2	3	GPCM	0.5754	-0.1764	
FR3	4	GPCM	0.7879	-0.0002	
FR4	4	GPCM	0.6943	-0.6524	
FR5	5	GPCM	0.4346	0.7731	
FR6	5	GPCM	0.5829	0.3850	

Table A32 Equated IRT Item Parameters, NEAT, Design 2, Automated Raters (Reduced Noise), Form 2

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC1	2	3PLM	1.1229	-1.4904	0.1930
MC2	2	3PLM	0.8583	1.8624	0.2913
MC3	2	3PLM	0.6796	-1.1023	0.2437
MC4	2	3PLM	0.7485	-0.4656	0.2044
MC5	2	3PLM	1.1406	-1.0525	0.1255
MC6	2	3PLM	1.2543	0.8889	0.2633
MC7	2	3PLM	0.9716	-0.3367	0.0369
MC8	2	3PLM	0.8303	-0.9890	0.2260
MC9	2	3PLM	0.8461	1.0616	0.0606
MC10	2	3PLM	0.8349	0.5865	0.2155
MC11	2	3PLM	0.9303	1.3282	0.2418
MC12	2	3PLM	1.0150	-1.4748	0.1077
MC13	2	3PLM	0.8272	-1.7850	0.1966
MC14	2	3PLM	0.7862	1.8781	0.0421
MC15	2	3PLM	1.2145	-2.6534	0.2336
MC16	2	3PLM	0.5559	-1.2789	0.1363
MC17	2	3PLM	1.0180	-1.9091	0.1482
MC18	2	3PLM	0.7154	-1.8132	0.1527
MC19	2	3PLM	0.6712	-2.2975	0.2042
MC20	2	3PLM	1.0618	-0.9503	0.2398
MC21	2	3PLM	1.3734	-0.5955	0.0414
MC22	2	3PLM	0.8647	-0.4968	0.1769
MC23	2	3PLM	0.7184	1.5680	0.1757
MC24	2	3PLM	1.1645	-1.4494	0.2097
MC25	2	3PLM	0.9682	0.5501	0.1459
MC26	2	3PLM	0.8996	2.1013	0.2055
MC27	2	3PLM	1.7723	0.0845	0.2655
MC28	2	3PLM	1.1236	1.2015	0.0069
MC29	2	3PLM	0.8730	0.3957	0.2811
MC30	2	3PLM	0.7678	2.0379	0.0448
MC31	2	3PLM	0.7671	3.0729	0.1187
MC32	2	3PLM	0.8476	1.9810	0.2892
MC33	2	3PLM	0.6188	-0.3393	0.2239
MC34	2	3PLM	1.2132	-0.8001	0.1643
MC35	2	3PLM	0.8051	-0.1373	0.0509
MC36	2	3PLM	0.9259	2.1144	0.1267
MC37	2	3PLM	0.5687	0.3237	0.1637
MC38	2	3PLM	0.7883	-0.6525	0.2444
MC39	2	3PLM	1.1368	0.7297	0.0783

Item Type	Score Levels	IRT Model	a-Parameter	b-Parameter	c-Parameter
MC40	2	3PLM	1.1374	0.6032	0.2646
MC41	2	3PLM	1.0534	-0.7357	0.1988
MC42	2	3PLM	0.9495	-1.1945	0.1978
MC43	2	3PLM	0.8469	-3.0794	0.2028
MC44	2	3PLM	0.5323	1.1766	0.0558
MC45	2	3PLM	0.9725	1.2371	0.1978
MC46	2	3PLM	1.1506	-0.0971	0.0302
MC47	2	3PLM	0.8870	0.0614	0.0883
MC48	2	3PLM	0.8357	-2.1798	0.1109
MC49	2	3PLM	0.9017	1.1061	0.0794
MC50	2	3PLM	0.8455	-0.3367	0.2112
MC51	2	3PLM	0.8981	1.1485	0.0738
MC52	2	3PLM	1.0066	1.4389	0.0000
MC53	2	3PLM	0.7636	0.9521	0.2113
MC54	2	3PLM	1.4081	0.5403	0.1005
FR1	3	GPCM	1.0886	1.1554	
FR2	3	GPCM	0.7671	-0.3039	
FR3	4	GPCM	1.2483	-0.1263	
FR4	4	GPCM	0.9258	-0.7680	
FR5	5	GPCM	0.5402	0.7448	
FR6	5	GPCM	0.6874	0.3977	

REFERENCES

Ajay, H.B., Tillett, P.I., & Page, E.B. (1973). *Analysis of essays by computer (AEC-II)*. Final 58 Report to the National Center for Educational Research and Development for Project, (8-0101).

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole

Attali, Y., & Burstein, J. (2006). Automated essay scoring with *e*-rater v.2. *Journal of Technology, Learning, and Assessment*, *4*, 1-30.

Barrett, M.D., & van der Linden, W.J. (2017). Optimal linking design for response model Parameters. Journal of Educational Measurement, 54(3) 285-304.

Bennett, R. E. & Behar, I. I. (1998). Validity and automated scoring: It's not only in the scoring. *Educational Measurement: Issues and Practice*, *4,* 9-17.

Bennett, R.E., & Zhang, M. (2016). Validity and automate scoring. In F. Drasgow (Ed.), Technology and testing: Improving educational and psychological measurement. New York: NY.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In, F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*. Read-ing, MA: Addison-Wesley

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. Journal of Educational Measurement, 25(4), 275-285.

Boyer, M., Kieftenbeld, V. (2016). *Evaluating automated rater rerformance: Is the state of the art improving*? Presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. New York: Academic.

Bernardin, H.J., LaShells, M.B., Smith, P.C., & Alvares, K.M. (1976). Behavioral expectation scales: Effects, of developmental, procedures and formats. *Journal of Applied Psychology,* 61, 75-79.

Casabianca, J.M., Junker, B.W. & Patz, R.J. (2016). Hierarchical rater models. In W.J. van der Linden (Ed.), Handbook of Item Response Theory Volume One: Models. Boca Raton, FL: CRC Press.

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

Cohen, J. (1968). Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Conley, D. T. (2014). *The common core state standards: Insight into their development and purpose* (Research Report). Retrieved from Council of Chief State School Officers' website: http://www.ccsso.org/Documents/2014/CCSS_Insight_Into_Development_2014.pdf

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Dorans, N.J. (1990). Equating methods and sampling designs. *Applied Measurement in Education,* 3(1) pp. 7-33.

Dorans, N. J. (2004). Using subpopulatoin invariance to assess test score equity. *Journal of Educational Measurement*, 4, 43-68.

Dorans, N. J., Pommerich, M., & Holland, P.W. (Eds.). (2007). *Linking and aligning scores and scales*. New York, NY: Springer-Verlag.

Eignor, D.R., Stocking, M.L., & Cook, L.L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education,* 3(1) pp. 37-52.

Efron, B., & Tibshirani, R.J. (1993). An introduction to the bootstrap (Monographs on Statistics and Applied Probability 57). New York: Chapman & Hall.

Garcia, S. & Herrara, F. (2008). An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, *9*, 2677-2694.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22, pp. 144-149.

Haberman, S. & Dorans, N.J. (2009). *Scale consistency, drift, stability: definitions, distinctions and Principles*. Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME), San Diego, CA

Hagge, S. L. & Kolen, M. J. (2012). Effects of group differences on equating using operational pseudo-tests. In M. J. Kolen & W. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating (volume2)*. (CASMA Monograph Number 2.2) (pp.45-86). Iowa City, IA: CASMA, The University of Iowa.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Han, K. T. (2009). IRTEQ: Windows application that implements IRT scaling and equating [computer program]. *Applied Psychological Measurement*, 33(6), 491-493.

Higgins, D., & Heilman, M. (2014). Managing What We Can Measure: Quantifying the Susceptibility of Automated Scoring Systems to Gaming Behavior. *Educational Measurement: Issues and Practice, 33(3), 36-46.*

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.

Kamata, A., & Tate, R. L. (2005). The performance of a method for the long-term equating of mixed format assessment. *Journal of Education Measurement*, 42, 193-213.

Keller, L.A, & Keller, R.R. (2013). The long-term sustainability of IRT scaling methods in mixed-format tests. *Journal of Educational Measurement*, 50(4) pp. 390–407.

Keller, L.A, & Keller, R.R. (2011). The Long-Term Sustainability of Different Item Response Theory Scaling Methods. Educational and Psychological Measurement, 71(2) pp. 362–379.

Kieftenbeld, V., Boyer, M. (2017). Statistically Comparing the Performance of Multiple Automated Raters across Multiple Items. *Applied Measurement in Education*, 30(2) pp. 117-128.

Kim, S., & Kolen, M.J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic cure methods. *Journal of Educational and Behavioral Statistics*, 32, 371-397.

Kim, S. & Walker, M. E. (2009). *Evaluating subpopulation invariance of linking functions to determine the anchor composition of a mixed-format test* (Research Report 09-36). Princeton, NJ: Educational Testing Service.

Kim, S. Walker, M. E. & McHale, F. (2010a). Comparisons among designs for equating mixed format tests in large-scale assessments. *Journal of Educational Measurement*, 47, 36-53.

Kim, S. Walker, M. E. & McHale, F. (2010a). Investigating the effectiveness of equating designs for constructed-response test in large-scale assessments. *Journal of Educational Measurement*, 47, 186-201.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adoptive testing. *Applied Psychological Measurement*, 8, 147-154.

Kneeland, N. (1920). That lenient tendency in rating. *Personnel Journal*, 7, 356-366.

Kolen, M.J. & Brennan, R.L. (2004). *Test equating, scaling, and linking: methods and practices, 2nd edition*. New York, NY: Springer-Verlag.

Kolen, M.J. & Brennan, R.L. (2014). *Test equating, scaling, and linking: methods and practices, 3rd edition*. New York, NY: Springer-Verlag.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M.D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Hillsdale, NJ: Lawrence Erlbaum Associates.

Landy, F. J., & Trumbo, D. A. (1976). *The psychology of work behavior.* Homewood, IL: Dorsey Press.

Leacock, C, & Chodorow, M. (2003). *c*-Rater: automated scoring of short-answer questions. *Computers and the Humanities, 37*, 389-405.

Leacock, Claudia, Gonzalez, Erin, Conarroe, Mike. (2014). Developing effective scoring rubrics for AI short answer scoring. McGraw-Hill Education CTB Innovative Research and Development Grant. Monterey: McGraw-Hill Education CTB.

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

Liu, O.L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., Linn, M.C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practices*, 33(2) 19-28.

Liu, O.L., Rios, J.A., Heilman, M., Gerard, L., & Linn, M.C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching, 53(2) 215-233.*

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F.M. (1982). Standard error of an equating by item response theory. *Applied Psychological Measurement*, 6 463-472.

Loyd, B. H. and Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 17 (3), 179-193.

Manning, C. D., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 14 (2), 139-160.

McCormick, E. J., & Tiffin, J. (1974). *Industrial psychology* (6th ed.). Englewood Cliffs, N. J.: Prentice-Hall.

Miller, M.D., & Linn, R.L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, 25(3) 205-219.

Morris, G.M. (1982). On the foundations of test equating. In P.W. Holland & D.B Rubin (Eds.), Test Equating. New York, NY: Academic Press.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data* [computer program]. Chicago, IL: Scientific Software.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review*, 51, 1-23.

Ogasawara, H. (2001). Standard errors of item response tehory equating/linking by response function methods. *Applied Psychological Measurement,* 25, 53-67.

Page, E.B. (1966).The imminence of grading essays by computer. *Phi Delta Kappan*, 48, 238–243.

Patz, R.J., Junker, B.W., Johnson, M.S., Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27(4), 341-384.

Perelman, L. C. (2013). Critique of Mark D. Shermis & Ben Hamner, "Contrasting state-of-the-art automated scoring of essays: Analysis." *Journal of Writing Assessment*, Retrieved from: http://journalofwritingassessment.org/article.php?article=69

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramineni, C., & Williamson, D.M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18, 25–39.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrica, 70*(1), 41–55.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetricTM essay scoring system. *The Journal of Technology, Learning and Assessment*, *4*, 1–21.

Saal, F. E., & Landy, F. J. (1977). The mixed standard rating scale: An evaluation. *Organizational Behavior and Human Performance*, IS, 19-35.

Scott, W. (1955). "Reliability of content analysis: The case of nominal scale coding." Public Opinion Quarterly, 19(3), 321-325.

Sheehan, K.M., & Mislevy, R.J. (1988). Some consequences of the uncertainty in IRT linking procedures (RR-88-38-ONR). Princeton, NJ: Educational Testing Service.

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment*, *20*(1), 46–65.

Shermis, M. D., & Hamner, B. (2012). *Contrasting state-of-the-art automated scoring of essays: Analysis*. Presented at the National Council of Measurement in Education, Vancouver, BC, Canada.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art in machine scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation*. New York, NY: Routledge.

Skaggs, G., & Lissitz, R.W. (1986). IRT test equating: relevant issues and a review of recent research. *Review of Educational Research*. 56(4), 495-529

Stocking, M.L, & Lord. F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal Educational Measurement*, 36, 336-346.

Tate, R. L. (2000). Performance of a proposed method for linking of mixed format tests with constructed response and multiple choice items. *Journal Educational Measurement*, 37, 329-346.

Taylor, C.S., & Lee, Y. (2009). Stability of Rasch scales over time. *Applied Measurement in Education*, 23(1) 87-113.

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

van der Linden, W.J., & Barrett, M.D. (2016). Linking item response model parameters. *Psychometrika*, 81, 650-673.

Verhelst, N. D., & Verstralen, H. H. F. M. (2001). An IRT model for multiple raters. In A. Boomsma, M. A. J. Van Duijn, and T. A. B. Snijders (Eds.), *Essays on item response modeling*. New York, NY: Springer-Verlag

Way, W.D., Davis, L.L., & Fitzpatrick, S.J. (2006). *Score comparability of online and paper administrations of the Texas Assessment of Knowledge and Skills*. Paper presented at the annual meeting of National Council on Measurement in Education, San Francisco, CA.

Way, W.D., Lin, C., & Kong, J. (2008). *Maintaining score equivalence as tests transition online: Issues, approaches, and trends*. Paper presented at the annual meeting of the national council on measurement in education, New York, NY.

Way, W.D., Um, K., Lin, C., & McClarty, K.L. (2007). *An evaluation of a matched samples method for assessing the comparability of online and paper test performance*. Paper presented at the annual meeting of the national council on measurement in education, Chicago, IL.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. Assessing Writing, 6, 145–178.

Wesley, K. N., & Yukl, G. A. (1977). *Organizational behavior and personnel psychology.* Homewood, IL: Irwin.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*, 2–13.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, 26, 283-306.

Wong, C.C. (2015). Asymptotic standard error for item response theory true score equating of polytomous items. *Journal of Educational Measurement*, 52(1) 106-120.

Yang, Y., Buckendahl, C. W., J., Piotr J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, 15(4), 391-412.

Yang, Y., Buckendahl, C. W., Juszkieweicz, P. J., & Bhola, D. S. (2005). Evaluating computer automated scoring: Issues, methods, and an empirical illustration. *Journal of Applied Testing Technology*, 7(3).

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245-262.

Yen, W.M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, 17, 297-311.

Yu, L., Livingston, S.A., Larkin, K.C., & Bonnett, J. (2004). Investigation difference in examinee performance between computer-based and handwritten essays (RR-04-18). Princeton, NJ: Educational Testing Service.

Zedeck, S., & Blood, M. R. (1974). *Foundations of behavioral science research in organizations.* Monterey, CA: Brooks/Cole.

Zeng, L., & Kolen, M.J. (1994) IRT scale transformations using numerical integration. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, 103, 374-378.