

Distribution Agreement

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

Signature:

A. Adrienne Walker

Date

Validity, Model-Data Fit, and Person Response Functions in Educational Assessment

By

A. Adrienne Walker
Doctor of Philosophy

Educational Studies

George Engelhard, Jr.
Advisor

Yuk Fai Cheong
Committee Member

Robert J. Jensen
Committee Member

Accepted:

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

Date

Validity, Model-Data Fit, and Person Response Functions in Educational Assessment

By

A. Adrienne Walker
M.A., University of Massachusetts, 2001
B.A., Agnes Scott College, 1997

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

An abstract of
a dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in
Educational Studies
2016

Abstract

Validity, Model-Data Fit, and Person Response Functions in Educational Assessment By A. Adrienne Walker

Background: A test score alone is not sufficient to represent a person's level of knowledge and skills on a measured construct. Validity of the interpretation and use of test scores is based on an underlying theory (APA/AERA/NCME, 2014), and measurement models based on item response theory provide one way to evaluate validity.

Good model-data fit at the group and individual levels is critical for establishing the validity of test score interpretation and use. Procedures for examining model-data fit have been developed and are used in educational testing. However in practice, these procedures are limited to ensuring adequate item-level fit and global person-level fit (i.e., person fit over all test takers). Procedures ensuring adequate individual person-level fit are not conducted for most educational tests. Furthermore, communicating person fit information to educational stakeholders who use test scores to make important educational decisions is also absent.

Purpose: This study explores an approach for examining and communicating individual person-level fit. The research questions addressed by the study are

1. How do person response functions and person-level model-data fit contribute to the validation of inferences regarding person scores?
2. What existing methods of creating person response functions can be utilized in practice for validating the inferences of scores on educational tests?

Methods: A review and critique of the literature provided the conceptual foundation for the study. I built upon this foundation by conducting three empirical applications that used real and simulated test data. A two-step, statistical and graphical, procedure was used to detect and illustrate individual person misfit. Specifically, person fit was examined statistically with person fit indices and visually with person response functions (PRF).

Findings: Individual person fit analyses and person response functions together have promise for inclusion in quality checking because they can illustrate test score trustworthiness in a clear way. Person response functions can be used as a tool to help researchers and practitioners understand individual person *misfit* in educational tests.

Significance: Individual person fit analyses provide information that validates the claims of test score meaning. This is a necessary, and currently missing, piece of validity information in large-scale educational testing practice.

Validity, Model-Data Fit, and Person Response Functions in Educational Assessment

By

A. Adrienne Walker
M.A., University of Massachusetts, 2001
B.A., Agnes Scott College, 1997

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

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in
Educational Studies
2016

Acknowledgements

The successful completion of this dissertation was made possible by a number of people. First, to my #1 man, my partner, my home: Billy Mallon, without your unwavering emotional and physical support, my pursuit of academic happiness would have fallen by the wayside. Thank you for being my rock on which to lean and to which to complain. Second, to my sweet and beautiful sons, Liam and Xander Mallon: Thank you for your patience and receptivity to mama going to school, mama being the teacher, and mama working on the computer all the time. This journey is not the most important one that I will do in my life (that honor is reserved for being your mom), but I believe it has shaped the most important journey for the better. I will endeavor to prove this to you.

The next three people who I wish to acknowledge have been with me since the beginning—my committee chair, Professor George Engelhard, Jr. and my committee members, Professor Fai Cheong and Professor Robert Jensen. For all three of you, let me start by saying thank you for always having my best interest in mind. Professor Engelhard, thank you for helping me to discover and cultivate my intellectual home in educational measurement and Rasch measurement theory. Aside from my family, I have no better cheerleader. And there is no one better than you at seeing the positive potential in someone rather than her current demonstrated ability. I am and will be eternally grateful for your mentorship, regarding scholarly pursuits, as well as non-scholarly ones. I will rely on this mentorship always.

Professor Cheong, I do not know how you manage to be both hard-nosed and patient at the same time, but I am very thankful for these *warm, demanding* characteristics of your teaching style. I always feel better prepared and more theoretically and practically grounded after discussing my work with you. You always know where to push my reasoning and how hard to push. Indeed, I do not know how I will manage new projects without your guidance and advice! Thank you for the countless hours you spent on making me a better researcher.

Professor Jensen, you have an uncanny way of seeing “the point” that I try to make straight away. This, along with your acumen for situating any research problem into a bigger picture, provides what all emergent scholars need: validation of the importance of their work. Your gift is that you make all of us (students) feel this way. Thank you for the practical advice and recommendations about my work and for validating the purpose of it.

To my mom and dad, Melva and Sloan Walker, I am so grateful for your emotional support that often manifested itself financially and logistically over these four (plus) years. Without additional money and childcare help, I would’ve lost my sanity a while ago! To my Grandma Walker and Grandma Buzbee (posthumously), thank you for encouraging me trust myself—in this case, to pursue more education just because I think it is best for me and my family...and for acting interested when I blather on about my research.

Lastly, but certainly not least, I'd like to thank my DES colleagues. To Dr. Avant, Simone and Robin, thank you for helping me navigate the administrative procedures of the Laney Graduate School. To Stefanie, Shanna, Adrienne, Morgan, Kris, Pati, Amber, and Miyoshi, thank you for making my time at Emory enjoyable and well-practiced. I have learned so much from hearing about your research pursuits. We did it!

Table of Contents

Chapter One: Introduction	1
Statement of the Problem	6
Purpose of the Study	7
General Research Questions.....	8
Theoretical Framework	9
Summary	16
Chapter Two: Review of the Literature.....	17
Validity: Meaning and Interpretation of Test Scores.....	17
Model-data Fit and Person Fit in the Rasch Model.....	28
Person Response Functions and Model-Data Fit	32
Uses of Person Response Functions.....	40
Summary	53
Chapter Three: Exploring Person Fit with an Approach Based on Multilevel Logistic Regression.....	57
Method	65
Results	71
Discussion	78
Chapter Four: Exploring Aberrant Responses Using Person Fit and Person Response Functions.....	83
Purpose	84
Method	93
Results.....	96
Discussion	101
Chapter Five: Using Person Fit Statistics and Person Response Functions to Validate Theta Estimates from Computer Adaptive Tests.....	105
Purpose	107
Method	114
Results.....	125
Discussion	132
Chapter Six: Discussion.....	137
References.....	148
Appendix A.....	185
Appendix B.....	186
Appendix C.....	187

List of Tables

Table 1. Summary of Person Response Functions Research (1941 to 2014)	166
Table 2. Parameter Estimates for the Multilevel Logistic Models	168
Table 3. Summary of Findings from the Anchored Rasch Calibration	169
Table 4. Mean Person Fit Statistic Values at the 95th Percentile	170
Table 5. Psychometric Information for 25 Examinees	171

List of Figures

Figure 1. Example Expected and Observed Person Response Function.	172
Figure 2. Conceptual Outline.....	173
Figure 3. Item and Person Response Functions for the Rasch and 3-PL.....	174
Figure 4. Rasch Infit and Outfit Fit by MLR Estimates.	175
Figure 5. Estimated person response functions and residuals.	176
Figure 6. Variable Map for the Anchored Rasch Analysis.....	177
Figure 7. Fitting person response functions.....	178
Figure 8. Misfitting person response functions, blind guessing	179
Figure 9. Misfitting person response functions, other guessing.....	180
Figure 10. Conceptual framework of person fit in CAT.....	181
Figure 11. Person response functions illustrating misfit by Outfit.....	182
Figure 12. Person response functions illustrating misfit by Infit.....	183
Figure 13. Person response functions illustrating misfit by Bfit.	184

Chapter One: Introduction

In educational settings, decisions about student knowledge and skills are made using achievement test scores. But how do educational researchers, practitioners, and policymakers know that a test score is a good representation of what a student knows and can do, and what the student should learn next? At the heart of this issue is the concept of validity or more specifically, the validity the inferences about student knowledge and skills that are made on the basis of test scores.

Validity is of paramount importance because it informs the meaning or interpretation of the score and consequently how trustworthy it is for its intended purpose (Messick, 1995). Results from educational tests must be justified for the use of describing a student's level of achievement in mathematics, language arts, or science. This justification is based on evidence that is accumulated throughout the entire test development and administration processes (Anastasi, 1986, 1988).

Validity is an ongoing collection of evidence that informs each step of the test development process (Messick, 1989, 1995). For example, validity informs the purpose or rationale for developing a test and establishes the boundaries for use of test scores. Validity informs what content and item formats should be included on the test in order to measure a construct adequately and accurately. Validity informs the selection of the items that are chosen to measure students' achievement. Validity informs the trustworthiness of the test results for making inferences regarding student performance. And validity informs the consequences of using test scores for making educational policy, curricular, or academic decisions about a student or groups of students.

In this dissertation, one aspect of validity evidence is examined: *measurement validity*, or from the list of examples above, *the trustworthiness of the test results for making inferences regarding student performance*. Zumbo (2007, 2009) uses the term measurement validity to refer to the set of assumptions that must be tested to validate the use of a psychometric tool. For this dissertation, measurement validity refers to the adequacy of a mathematical model (and the test scores derived from it) for predicting students' responses to the test items. The extent that the mathematical model can predict student responses to the items is the extent to which measurement validity is demonstrated and credible measures of student achievement are obtained by a student's test performance. This piece of validity evidence is necessary for validating the trustworthiness of a score for making inferences about a student. This general idea is similar to the step of model testing where a hypothesized model is applied to new set of data and the predictive strength is evaluated.

Today, the psychometric models that are used to guide test development are predominantly derived from item response theory (IRT). For tests designed with IRT, measurement validity can be established by evaluating the responses that students give to the individual items that make up the test. This evaluation is often undertaken by using model-data fit procedures (Embretson & Reise, 2000; Swaminathan, Hambleton, & Rogers, 2007). Model-data fit describes how well the responses that students give to the items match with an expected pattern of responses that is based on a theoretical model of this relationship.

As with all models, there is rarely perfect alignment between what is theorized and what is observed, yet when a close match between observed responses and the

theoretical model is realized, the model can be considered an adequate representation of the relationship between the achievement level of the student and the characteristics of the items. This result provides one piece of validity evidence that the test scores can be considered adequate representations of student knowledge, within the context of the designed purpose and scope of the test.

One way to evaluate model-data fit is by conducting model comparisons.

Measurement models that vary in complexity can be compared, and the model that has the smallest amount of residuals, the differences between what response is given and what response is expected based on the measurement model for a particular person or item, would be chosen to represent the data. When one measurement model is preferred over other measurement models because of the mathematical properties of the model or because of how a test is designed, the actual test data may be constrained or altered until an adequate level of model-data fit is achieved.

IRT models have been shown to be robust to imperfect model-data fit, but model-data fit should always be examined (Hambleton, Swaminathan, & Rogers, 1991; Swaminathan, Hambleton, & Rogers, 2007). Because model-data fit is continuous (it ranges from very poor to very good), thresholds or guidelines are often used in practice to signify when model-data fit is “good enough” to provide useful and valid measures of student achievement. Three main categories of model-data fit analyses are those that evaluate unidimensionality, item fit, and person fit. The categories of unidimensionality and item fit are only briefly introduced here because they are not the focus of this research. More information about these categories and their corollaries can be obtained from Hambleton et al. (1991), Meijer and Sijstma (2001), and Swaminathan et al. (2007).

Unidimensionality is the extent to which the items that make up a test measure one underlying trait. It is a primary assumption of traditional Item Response Theory (IRT) models.¹ Unidimensionality is important because when it holds for a set of item response data, it can be said that one trait, like mathematics achievement, describes students' responses to the items.

Item fit is the extent to which items accurately discriminate between persons who have different achievement levels. In other words, item fit describes whether or not the items have consistent difficulty levels within and across populations of students. Person fit can be described as the extent to which persons accurately discriminate between items with different difficulty levels, or in other words, if persons have consistent achievement levels regardless of which items they answer. Item and person fit are important for measurement because they provide the justification that the measuring instrument (e.g., the test items) is stable across the different persons who will be measured by them.

Item-fit procedures identify items that do not conform to the parameters established for them by an IRT model. For instance, an item that appears difficult when included on one test and easy when included on another test would be categorized as misfitting the measurement model. Items that cannot be placed reliably along the achievement continuum given the pattern of responses they elicit from examinees cannot contribute to the measurement of student achievement in a meaningful way. They are often revised (and re-evaluated) or discarded.

¹ Multidimensional IRT models (Reckase, 2009) have been developed that do not assume that a single dimension is measured by a test. However, these models have not been widely utilized in large-scale testing practice; therefore they are not discussed here.

In a similar fashion, person-fit procedures identify persons that do not respond to the items as expected given their total score and based on a criterion, such as what is expected by an IRT model or what is likely based on the set of response patterns observed in the data. For instance, a person with an average total score who answers many difficult items correctly and many easy items incorrectly would exemplify poor person fit. Persons who provide unexpected response patterns like this cannot be reliably placed along the achievement continuum. As a result, they are not measured by the test in a meaningful way.

Unlike misfitting items, misfitting persons cannot be revised or discarded, and at the end of the testing process, it is the person who is judged by his or her performance on the test. Poor person fit has implications for the interpretation and use of individual scores. The individual test scores of misfitting persons, if left alone, yield unsupportable inferences regarding their individual levels of achievement—their true achievement level may be over or underestimated. Moreover, because the psychometric characteristics of test scores for misfitting persons may not be the same as the psychometric characteristics of test scores for fitting persons, comparing scores for these two groups of persons is problematic (Meijer, 2002).

Over the past two decades, researchers' rationales for studying person-fit has expanded from that of evaluating overall (or global) model-data fit to evaluating model-data fit of individuals or small groups of students (for example, Engelhard, 2009; Lamprinou, 2010; 2013; Lamprinou & Boyle, 2004; Perkins, Quaynor, & Engelhard, 2011, Petridou & Williams, 2007, 2010). In this dissertation, I continue the expansion of examining person fit at the individual level. My rationale is that this information is *vital*

for the appropriate use of individual test scores. Moreover, I argue that communicating information about person fit to general test score users can bring issues of validity to the forefront of mainstream test score use.

In this dissertation, I explore ways of detecting and conveying person-misfit that can be easily implemented at local educational agency levels with respect to the requirements for technical expertise or computer software. I believe these are the initial steps necessary for introducing the idea of person fit into test reporting practice. I argue that by utilizing statistical and graphical methods for examining person-fit, researchers can make person fit information useful and accessible, not just to other researchers, but to all test score users. It is my plan that this research will promote awareness that all test scores are not equally trustworthy for representing student knowledge.

Statement of the Problem

In current testing practice, model-data fit and other quality assurance procedures are used to ensure the necessary psychometric and statistical properties are met by the test items. Because of advances in item writing and test development processes, adequate levels of model-data fit are usually observed for the whole set of test data. Adequate model-data fit indicates that, in general, the test scores can help show what the group of students knows, what they can do, and can help inform what they should learn next. This provides one piece of validity evidence.

But despite acceptable overall model-data fit, it is still possible for some students to provide responses patterns that do not fit the model adequately. For these students, the test scores may not be good representations of what they know, can do, and what they should learn next. In other words, not all test scores are equally trustworthy

representations of student knowledge and skills. Given this variation, it seems reasonable that the examination of person fit at the individual level should be conducted in order to promote the appropriate interpretation and use of individual test scores. Moreover, this information should be provided to practitioners and other test score users.

Procedures for examining person fit have been developed and refined (Karabotsos, 2003; Meijer & Sijtsma, 2001). Some procedures that explain and visually illustrate person fit have also been developed (Emons, Sijtsma, & Meijer, 2004, 2005; Reise, 2000; Strandmark & Linn, 1987). But despite these developments, procedures for examining *individual* person fit are not used in testing practice (Cui & Roberts, 2013). Moreover, the problem surrounding the trustworthiness of a test score from a *misfitting* response pattern appears to be known only among psychometricians, not the general educational stakeholder population. The absence of the examination and reporting of individual person fit represents a gap between testing research and practice. In this dissertation, I explore a potential way to fill this gap.

Purpose of the Study

The purpose of the study is to explore the usefulness of person response functions as an approach for examining, evaluating, and communicating person misfit. Person response functions are graphical representations of the relationship between the probability of a person giving a response and the difficulty level of the items that make up a test. Because they are represented visually, person response functions may be a promising way to convey information about person fit to researchers, educational stakeholders, and practitioners.

Previous researchers have explored methods for creating person response functions (Carroll, Mead, & Johnson, 1991; Emons et al., 2005; Engelhard, 2013a; Reise, 2000; Sijstma & Meijer, 2001). Much of this research has been conducted using simulated data, although some researchers have applied the methods to real educational or psychological test data. Although most previous research links person response functions to measurement validity and to appropriate test score inferences, none of the aims of the previous research studies that I have read, mention explicitly the practical concern of using person response functions as a way to inform test score trustworthiness and to enhance and encourage appropriate test score use. This study extends previous work using person response functions by exploring several ways to create person response functions with educational data. Furthermore, the methods used to create the person response functions in this study were chosen based on feasibility and evaluated based on interpretational clarity for conveying person fit information to a general educational stakeholder audience.

General Research Questions

This research is guided by the following research questions:

1. How do person response functions and person-level model-data fit contribute to the validation of inferences regarding person scores?
2. What existing methods of creating person response functions can be utilized in practice for understanding the patterns of person responses and validating the inferences of scores on educational tests?

The first question was answered by a review of the literature. The second question was answered by three empirical applications. These applications are *Exploring Person Fit*

with an Approach Based on Multilevel Logistic Regression, Exploring Aberrant Responses Using Person Fit and Person Response Functions, and Using Person Fit Statistics and Person Response Functions to Validate Theta Estimates from Computer Adaptive Tests. In these applications, approaches for creating person response functions were explored. Each application was designed as a stand-alone study with specific research questions.

Theoretical Framework

In current testing practice, there are several commonly used item response theory models. Although assessing model-data fit for these IRT models is conceptually similar, procedurally it is different. For this study, the Rasch measurement model is chosen as the item response theory model.

In the subsequent paragraphs, the theoretical framework for the study is introduced. First, a brief introduction of “measurement” from an item response theory perspective is provided. Next, Rasch measurement theory is introduced, followed by the introduction of the dichotomous Rasch model and the specific model-data fit procedures used to assess Rasch model person fit. Lastly, the connection between Rasch model person fit and person response functions is discussed.

Measurement in Item Response Theory

For tests that are designed with item response theory (IRT), the philosophy of measurement is based on the concept of an underlying latent variable that is believed to be related to a person’s performance on a series of test items. Levels of the underlying variable can be thought of as existing on a continuum, which ranges from less of the variable to more of the variable. The ultimate goal of an IRT model is to transform a

person's responses to the items into a location along the latent variable continuum that describes how much of the latent variable he or she possesses.

A major benefit of using IRT models is that the locations, or measures, of persons along the latent variable do not depend on the set of test items that the persons answered. That is, person measures of achievement are comparable if they were administered a set of easy test items or if they were administered a set of difficult test items. This benefit is unique to item response theory because during the analysis process, the items are also placed along the latent variable and these item calibrations exist in the same metric as the person measures (Baker, 2001).

IRT models assume that a person with a higher level of achievement have a higher probability of giving the correct answer to an item than a person with a lower level of achievement. The exact mathematical form and the number of item and person characteristics it takes to model the persons responses to the items differs across item response theory models, but all IRT models have at least one parameter for items and one parameter for persons. Moreover, adequate fit between the IRT model and the response data being analyzed is needed to ensure that the ideal definition of measurement properties is attained in practice (Swaminathan et al., 2007). For these reasons, model-data fit analyses are a critical part of any application of IRT.

Invariant Measurement and Rasch Measurement Theory

Although the properties of measurement invariance hold for all IRT models, some researchers in the IRT community argue that *invariant measurement* is only achieved when both the persons and the items have stable relative ordering of difficulty along the latent variable (Andrich, 2004; Anderson, 1973; Bond & Fox, 2007; Engelhard, 2013b;

Wright, 1992; Wright & Panchapakesan, 1969). Thus, the difference between invariant measurement and measurement invariance is that the order of difficulty for a set of test items must hold for all achievement levels along the latent variable *and* that the order of achievement level of persons must hold across all levels of item difficulty. For measurement invariance, the order of persons must be invariant across the items, but the difficulty of items may be different for different levels of person achievement.

The additional restriction is sometimes referred to as invariant item ordering (IIO, Sijtsma & Molenaar, 2002) and in an IRT framework, this requirement is unique to Rasch measurement models (Rasch 1960/1980) or models that have similar restrictions for item parameters (Sijtsma & Hemker, 1998).² A core set of requirements for invariant measurement in the Rasch model has been presented by Engelhard (2013b, pp.13-14).

These are:

1. The measurement of persons must be independent of the particular items that happen to be used for the measuring.
2. A more able person must always have a better chance of success on an item than a less able person.
3. The calibration of the items must be independent of the particular persons used for calibration.
4. Any person must have a better chance of success on an easy item than on a more difficult item.

² For instance in the non-parametric IRT framework (NIRT) a model exists that also requires invariant item ordering—the Double Monotonicity NIRT model (DM, Sijtsma & Molenaar, 2002). The DM model is very similar conceptually and empirically to the Rasch model (Engelhard, 2008; Meijer, Sijtsma, & Smid, 1990).

5. Items and persons must be simultaneously located on a single underlying latent variable.

The first two requirements are concerned with invariant person measurement, and the second two are concerned with invariant item calibration. The last requirement is concerned with the opportunity to show the item and person locations on the same continuum (that represents the latent variable) on what is called a variable map. The variable map requirement necessitates the invariant ordering of items across different levels of achievement because without invariant item ordering the set of items and persons could not be shown in a single latent variable space (Engelhard, 2013b).

The Rasch model has had an enduring existence in applied educational testing programs. Because of its strict requirements, the Rasch model yields measures that are intuitively interpretable by practitioners (Baker, 2001). Model-data *misfit* to the Rasch model can also be observed in a straightforward manner because of the strict requirements. For these two reasons, the Rasch model was selected for use in this study. Moreover, because the data for this research required dichotomous scoring (the response was either right or wrong), the Dichotomous Rasch model was employed.

Dichotomous Rasch Measurement Model

The dichotomous Rasch measurement model describes the response to an item as a function of the location of the person and the location of the item along the latent variable (Wright & Stone, 1979). Mathematically, the Rasch model is formulated as

$$\phi_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad [1]$$

where, ϕ_{ni} represents the probability of person n with a location θ_n (achievement level) on the latent variable giving the correct response (denoted as 1) to item i with a location δ_i (difficulty) on the latent variable.

As mentioned earlier, applying the Rasch model to a set of data is not enough to produce the properties of invariant measurement. Test data must fit or approximate the model before invariant measurement is achieved. In other words, the extent to which the model can successfully reproduce the item response data that are observed indicates if the model provides a viable framework for interpreting the response data.

Model-data fit procedures typically employed for checking the fit of a Rasch model to a set of data were used in this research, and the details of these procedures are provided in later chapters of this dissertation. What is important to note at this time is that making a judgment about whether or not good enough model-data fit has been observed is a subjective decision. In practice, the level of fit necessary for measurement depends on the particular testing context. For instance, for tests that have higher-stakes associated with the outcome, more stringent criteria for model-data fit may be required than for tests that have lower-stakes associated with the outcome.

It is also plausible (and likely) for a set of test data to meet the requirements of invariant measurement as a whole, but for some individuals within the set to not meet the requirements. Adequate global fit, but poor individual fit implies that the location of the items are adequately stable and that it is likely that only a few persons are responding to the items unexpectedly. Investigating the *variation* in person fit to the model is a way to ensure that a construct is being measured the same way for all persons. Conceptually, this idea resembles idiographic methods for scientific inquiry (Walker & Engelhard,

2014) in that the relationship present for the aggregate may or may not explain the relationship present for the individual (Molenaar, 2009). For example, moderate variation in person fit could suggest that the construct may not be well-defined for certain individuals or groups of persons.

Person response functions (PRF) provide a visual way to delve deeper into individual person fit and are the graphical approach that is explored in this research. They are introduced in the next section.

Person Response Functions

Person response functions are graphical representations of the relationship between the probability of a person giving a response and the difficulty level of the items that are included on a test. In this study, two forms of PRF are used to evaluate person fit: expected PRF, which are based on the measurement model, and observed PRF, which are based on a person's actual responses to the items. The match between the expected and observed PRF represents how well a person's responses fit the model.

Figure 1 shows an example with expected and observed person response functions. In Figure 1, the y-axis represents the probability of a person giving the correct answer to a dichotomous item. The x-axis represents the difficulty levels of the items on a test.

Panel A of Figure 1 shows an expected person response function based on the Rasch model. This function shows the probabilities of a particular person giving the correct response to items on a test *when the response data fit the Rasch model*. These expected probabilities are calculated by inserting the person's achievement level value and the item difficulty values into the Rasch model introduced in Equation 1. They

represent the ideal pattern of responses for the items for a person with the achievement level, which in Figure 1 is 0.00 logits. The decreasing curve of the expected function shows that as the difficulty levels of the items increase on the x-axis, the probability of giving the correct answer decreases.

Panel B of Figure 1 shows an observed person responses function. The plotted probabilities are a function of the actual scored responses given by a particular person. There are different ways that these actual probabilities can be computed, and these methods are mentioned in later chapters of this dissertation. However, the point that is important to mention now is that the purpose of creating observed PRF is to graphically represent the pattern of responses that lies beneath a person's given responses. Generally, the pattern underlying the response vector is obtained by grouping the items by difficulty, and then calculating or estimating a summary value (such as a mean or median) of the person's observed responses to the items in that group. These summary values are then plotted. For instance, in Panel B of Figure 1, the pattern underlying the response vector was computed by grouping the items into four exclusive categories by item difficulty, calculating the mean of the person's observed responses to the items in that group, and plotting the results.

Unlike the probabilities of the expected PRF, which are dictated by the IRT model, the probabilities of the observed PRF follow the data. Because of this characteristic difference between the two types of PRF, a comparison of them can visually show the extent to which a person's observed response pattern matches with an expected response pattern. It is this idea of a visual evaluation of person fit that is pursued in this research.

Summary

This research explores the validity of the interpretation and use of test scores through the lens of item response theory using model-data fit; specifically person fit to the Rasch model. It is noted that there are many ways to examine validity in educational testing and this study represents one way and provides one piece of validity evidence. A key premise of this research is that graphical renderings of person fit using person response functions can communicate information about person fit. It was planned that through an exploration of person response functions, a useful way to facilitate substantive interpretations of person fit and to communicate information about person fit to researchers, practitioners, and other educational stakeholders would be found.

In this chapter, I linked person fit with validity using an aspect of validity called measurement validity (Zumbo, 2007, 2009). I did this in order to clearly situate person fit in a validity framework and to differentiate it from other types of validity evidence (e.g., content, consequential, criterion, construct). In the next chapter and throughout the rest of this research, however, I use the term *validity* instead of measurement validity because in modern frameworks, validity is seen as a unitary concept that encompasses all aspects of evidence used to support the measurement of a construct (e.g., Messick, 1995; Sireci, 2009).

Chapter Two: Review of the Literature

There are many ways to study validity in educational testing. With each way, more evidence is accumulated to support the judgement that the test measures what it purports to measure and that the measures are good. In this dissertation, I focus on one way of studying validity. I examine validity through the lens of model-data fit, and more specifically through the examination of person fit and person response functions.

Figure 2 illustrates the conceptual path that is followed in this study, specifically how person response functions and person fit are used as validity evidence supporting the inferences from test scores. It serves as the graphical organizer for the important topics of the dissertation. Chapter One provides the narrative that links these three topics and summarizes the research problem. In this chapter, a deeper discussion of validity, model-data fit, and person response functions, is provided by summarizing the relevant literature. It is noted that the topics of validity and model-data fit can stand-alone as dissertation topics, but in this research, their connection to person response functions (through person fit) is the focus. To best argue this connection, Chapter Two is divided into five areas: (1) validity, (2) validity and model-data fit, (3) model-data fit and person fit (in the Rasch model), (4) person response functions and model-data fit, and (5) uses of person response functions.

Validity: Meaning and Interpretation of Test Scores

The concept of validity spans over all scientific disciplines. For those disciplines that seek to measure phenomenon that are not directly observable, validity is integral to supporting value, worth, and usefulness to an outcome measure, such as a score on an

achievement test. For almost 100 years, researchers in social sciences have developed and revised practices and theories for establishing validity (Shear & Zumbo, 2014).

These practices and theories are still being revised today.

In its most basic and original definition, validity exists if an instrument measures what it claims to measure (Buckingham, 1921). In the earliest validity studies, theorists used statistical relationships between test results and other measures of the construct, conducted with the new techniques of correlation and factor analysis, to establish validity (Anastasi, 1986, Sireci, 2009). These methods are similar to what we call criterion-related validity today. Importantly, in these early conceptions, validity was thought of as a property of the test.

Discontent with the practice and definition of validity emerged among researchers. The original definition was criticized for being incomplete and not useful in practice (e.g., Loevinger, 1957), especially in light of measuring constructs for which an acceptable criterion or criterion instrument had not yet been created. A different perspective of validity emerged to include procedures beyond test/criterion relationships. During this time, theorists developed multiple types of validity. Content, predictive, concurrent, and construct validity could be used separately to establish validity. The more of these types that an instrument could show, the more valid it was presumed to be. These types of validity were included in a report called *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA/AERA/NCME, 1954) which served as professional guidelines or standards for testing practice (Sireci, 2009). These guidelines were the precursor to the first version of the *Standards for Educational and Psychological Testing*.

The types of validity reflected in early validity theory still hold a place in contemporary views of validity. However, instead of being conceptualized as separate types of validity, they are conceptualized as types of validity *evidence*. This shift in the conceptualization of validity is often credited, at least in part, to the seminal papers by Cronbach and Meehl (1955) and Loevinger (1957).

Cronbach and Meehl formally argued that tests measure unobservable phenomenon (traits) and that the connection between the construct that is desired to be measured and the items or stimulus doing the measuring needed to be supported theoretically and empirically. They proposed the idea of nomological nets, a set of psychological laws and observed relationships that explained how test results should relate to the construct being measured. These nomological nets were empirically established and were used to help build the theory and understanding of the construct of interest. In Cronbach and Meehl's paper, construct validity was highlighted as being the most important type of validity for forming meaning of test scores.

Loevinger (1957) also argued the importance of construct validity. She writes that construct validity is the "whole of validity from a scientific point of view" (1957, p. 636), and she advocated for partitioning construct validity into three subcategories: substantive, which focused on the theory-based analysis of test content, structural, which focused on internal structure of the test, and external, which focused on relationships to other criteria and systematic sources of error.

Although these two sets of researchers pushed for validity as a unitary concept with construct validity at its center, this shift from validity as separate types to validity as a unitary concept moved slowly through the 1960s and 1970s. Sireci (2009) argues that

the shift from *separate to unitary* validity became explicit in the 1974 version of the *Standards*. Then, in the 1980s and 1990s, some validity theorists re-argued that validity as an interactive system of evidence concentrating on construct validity. Messick (1980, 1989) is often credited with leading this charge; he argued that the purpose of validity is to establish score meaning and the values of score use for the construct being measured. In Messick's view any evidence that is collected about the test is used to strengthen or weaken the current theory about how the construct exists. Other validity theorists supported the conception that practically "any information gathered in the process of developing or using a test is relevant to its validity" because it contributes to the understanding of what the test measures (Anastasi, 1986, p.3).

Yet because any and all evidence collected pertained to validity, the validation process seemed to many researchers to be a never-ending and overwhelming task. In an effort to make test validation practical and doable, Kane (1992, 2006) provided guidance for test validation practices for specific testing contexts. In his *argument-based* approach to validation, Kane suggested compiling the sources of validity evidence in order to make validity arguments for the appropriate interpretations and uses of test scores in a given context and for a given purpose. In this view, the process of validation uses two arguments. First, the interpretive argument lists the proposed interpretations and uses of test results by organizing the network of inferences and assumptions "leading from the observed performance to the conclusions and decisions based on the performances" (Kane, 2006, p. 23). Second, the validity argument is an evaluation of the interpretive arguments. The rationale for these validity arguments is found in the writing of Cronbach

(1988) who argued that it is not the test that needs validating, but the interpretations and uses of the test results.

Today, validity theory and how it is conceptualized and practiced is still debated. Many theorists hold to the idea that validity centers around establishing arguments for construct validity (Sireci, 2007). Among these validity theorists, there are differences of opinion about how much emphasis should be placed on different sources of validity evidence (e.g., Embretson, 1983, 2007; Zumbo, 2007, 2009). There are also some theorists that disagree that construct validity is the center of validity educational and psychological tests (e.g. Borsboom, 2005; Lissitz & Samuelson, 2007). For instance, Borsboom (2005) argues that a psychological theory of response behavior that includes how possessing varying amounts of the construct should explain variations in the cognitive processes used to respond to items as well as in the outcome measures is necessary for establishing validity. Lissitz and Samuelson (2007) argue that content validity and reliability are the most important sources of validity evidence for educational tests and that construct validity need not be the central principle in these test settings.

In this dissertation, a definition and conceptualization of validity that is consistent with the current version of the *The Standards for Educational and Psychological Testing* (APA/AERA/NCME, 2014), the professional guidelines for the educational and psychological testing practice, is used. In the *Standards*, validity is defined as the “degree to which evidence and theory support the interpretation of test scores for proposed uses” (APA/AERA/NCME, 2014, p. 11). The conceptualization of validity used in this study, is characterized by five statements from Chan (2014, p. 10) and summarized here:

1. First, validity is not a property of the instrument, but instead it is about the inferences, claims, or decisions that are made based on the scores.
2. Second, construct validity is the focus—sources of validity evidence are accumulated and synthesized to support construct validity.
3. Third, the process of establishing validity is continuous and ongoing.
4. Fourth, in addition to the more traditional sources of validity (e.g. content and criterion), evidence based on response processes and consequences of test use and misuse should also be included in validation practice.
5. Fifth, there is no single way to validate test score meaning and use; the emphasis of validation practices such as establishing a validity argument, an explaining score variation, and item response process modeling can change depending on the context and purpose of the test (Chan, 2014, p. 10).

The Standards provide guidelines for the types of research that can be conducted to evaluate validity in educational tests. Six types of validity evidence are listed to help practitioners validate the intended interpretation and uses of test scores: Evidence based on content, response processes, internal structure, relations to other variables, and consequences of testing. In this research, validity evidence based on *response processes* is the primary focus.

Validity evidence based on response processes usually comes from analyzing individual responses. In this research, analyzing individual responses with person fit procedures is explored, and measurement theory provides the statistical framework for describing the test responses. Specifically, measurement theory describes how test performance should differ across persons who have different levels of achievement.

When test responses can be governed by measurement theory, evidence for the test measuring the construct (in a valid way) is obtained. When test responses cannot be governed by measurement theory, the claims about the test measuring the construct in a valid way are not supported. This means that something about the theory or the test is awry and must be revised before valid meaning can be attributed to the scores.

The most dominant measurement theory used in practice today is item response theory (IRT). Item response theory proposes that person measures on a construct of interest can be obtained by mathematically modeling the relationship between person characteristics and item characteristics. There are many different item response theory models that can be applied to a set of response data to obtain person measures of a construct, but the validity of any IRT model for representing item responses is examined and ultimately judged by using model-data fitting procedures (van der Linden & Hambleton, 1997). Adequate model data fit is critical for measurement; thus it is critical for establishing validity.

Validity and Model-Data Fit

Person fit represents one method out of several other methods for evaluating model-data fit. Hambleton, Swaminathan, & Rogers (1991, pp. 56-58) organize procedures for assessing model-data fit into three categories: *Checking Model Assumptions*, *Checking the Features of the IRT Model*, and *Checking Model Predictions*. For the discussion of model-data fit provided here, I borrow this organization because with it person fit procedures can be cleanly situated within the confines of an overall framework of model-data fit. It is beyond the scope of this dissertation to provide lengthy discussions of each aspect of model-data fit and the possible procedures used to

evaluate them. Instead, a brief discussion of main aspects of each of the three general categories of IRT model-data fitting procedures is provided. Then, the discussion segues into a general discussion of person fit and the way that person fit was inspected in this study.

The first category of IRT model-data fitting procedures as laid out by Hambleton et al. (1991) is *checking model assumptions*. The most commonly used IRT models today rest on assumptions of unidimensionality and local independence (Embretson & Reise, 2000). Unidimensionality describes the condition where only one latent variable can explain the shared variance across the item responses. When unidimensionality is present, a test score is a clear indicator of a single construct. Multidimensionality is problematic for measurement because when it is present, a test score becomes an ambiguous indicator of two or more constructs.

Local independence is a related concept in that it means that after controlling for a person's amount of the latent variable, the item responses are statistically independent of each other. With local independence, each item response can be conceptualized as adding a piece of information to the estimation of person's location along the latent variable continuum. When item responses are dependent on each other (locally dependent), the information gleaned from each item response is not *new* or *additional*, thus it cannot be assumed to add any additional information to the estimation of the person's location along the latent variable continuum.

Additional model assumptions exist, but are specific to the particular IRT model that is chosen. For instance, the Rasch IRT model also assumes that items have equal discrimination indices, and that targeted items minimize guessing. Goodness of fit tests

for these assumptions can be conducted using factor analytic, regression, residual, or other statistical procedures.

The second category for assessing model-data fit laid out by Hambleton et al. (1991) is *checking the features* of an IRT model. In other words, it is important to check whether or not the desirable features of measurement that are promised by the IRT model, exist after the model has been applied to the data. This category differs from the previous category in that it describes goodness of fit for particular IRT models instead of describing general fit for all IRT models.

The most desirable feature of any IRT model is the feature of invariance. Because IRT rests on the assumption that responses to items are due to the interaction between item and person characteristics, invariance should exist at the person and item level (de Ayala, 2009; Wright, 1968). Person invariance describes the condition where a person's performance on one set of items will yield the same location along the latent continuum (within measurement error) as his or her performance on a different set of items. Item invariance describes the condition where an item's performance for one set of persons will yield the same location along the latent continuum (within measurement error) as its performance for a different set of persons. Model-data fit for invariance can be evaluated by comparing the measures of persons (or items) from two groups with correlational analyses or scatter plots.

The third category of model-data fit assessment is *checking model predictions* (Hambleton et al., 1991). A measurement model represents a theoretical relationship. When the model fits the test data adequately, the model should be able to predict with some accuracy person's responses to the items. As with all statistical models, the fit to

real data will not be perfect, but gross violations of perfect prediction can imply poor model fit and the need for a different model to represent the data or different data to match the model.

Checking model predictions is typically conducted via examination of residuals. A residual is the difference between the actual item response and the item response that is expected based on the IRT model. A large negative or positive residual suggests that the expected response does not predict the actual response that was given. The more large residuals a person's response pattern has, the less accurate the model prediction is, and the poorer the model-data fit.

Although using residuals is conceptually straightforward, a limitation of using them is that it is difficult to interpret the degree of unexpectedness the responses show (Wright & Stone, 1979). It is difficult to judge what a normal amount of unexpectedness is and what an abnormal amount of unexpectedness is. And furthermore it is difficult to judge the point at which too many large residuals have been observed and the validity of the person measure (i.e., her score) becomes compromised? Standardizing the residuals helps researchers make judgements about the degree of unexpectedness of the response pattern because standardized residuals can be interpreted using a statistical framework for probabilistic events.

For the standardized residuals, the expected response and the binomial standard deviation of the expected response is taken into account. The residual is divided by the standard deviation of the expected value, and the standard deviation becomes the frame of reference to interpret the residual. Residual differences that are more extreme than the standard deviation are larger in absolute value, whereas residual differences that are less

extreme than the standard deviation are smaller in absolute value. Model data fit can be examined by inspecting the shape and spread of the standardized residuals. When the data fit the model, the standardized residuals should be distributed normally about 0.

In order to make a judgment about how well an IRT model represents a set of test data, other techniques that evaluate the standardized residuals at the overall model, item, and person levels are used. One way that can be used to examine overall model fit is through a model comparison approach (de Ayala, 2009; Embretson & Reise, 2000). Generally speaking, the goal of such an analysis is to choose the most parsimonious model that can still represent the data well. A complex model is compared to a less complex model using statistical procedures like the likelihood ratio or chi-square tests. If no significant difference exists between the fit of both models, then the reduced model (more parsimonious) can be selected as fitting equally well.

Item and person level fit can be assessed with statistical and graphical techniques. Statistical techniques test the significance of the residuals for each item or for each person. Individual persons or items that reveal misfitting response pattern can be reviewed and then removed or otherwise handled. Many different item and person fit statistics have been developed for detecting misfitting item and person responses. Like the fit statistics that assess the fit of the overall model, many of the individual item and person fit statistics are based on chi-square or log likelihood procedures. One criticism of these techniques is their sensitivity to sample size and their unknown sampling distribution under the null hypothesis that the item or person responses fit the model well.

Graphical techniques compare the expected item or person response functions with functions that are computed from the actual data. A benefit of such techniques is

that they can reveal areas along the latent variable continuum where misfit may be present. These residuals may implicate one or more possible reasons for the misfit. From the perspective of item misfit, both statistical and graphical indices of fit can assist test developers in creating better items to measure the intended construct. From the perspective of person misfit, both statistical and graphical indices are important because aberrant person responses can affect the item calibrations, which consequently affect person measures.

But the importance of person-fit extends beyond its importance to overall model-data fit. Examining person fit for the sole purpose of identifying persons whose responses to test items may be too haphazard or too perfect for their scores to be considered trustworthy, is important in its own right. Embretson and Reise (2000) write that “in a way, person-fit indices attempt to assess the validity of the IRT measurement model at the individual level and the meaningfulness of a test score derived from the IRT model” (p. 238). Person fit as a way to inform test score meaning and use for a specific individual is a research topic that has seen a rise in interest over the past two decades. The next section provides the rationale and background for the person fit techniques that were used in this dissertation.

Model-data Fit and Person Fit in the Rasch Model

When a set of test data fit the Rasch model closely, invariant measurement has been achieved and the estimates of the persons’ achievement levels and the items’ difficulty levels can be considered as trustworthy representations of the persons’ and items’ locations along the latent continuum. Yet, as was alluded to in the previous section, simply applying the Rasch model to a set of data is not enough to produce the properties

of invariant measurement. Model-data fit must be explicitly checked. According to Rasch (1960/1980), “*Models should not be true, but it is important that they are applicable*, and whether they are applicable for any given purpose must of course be investigated” (p. 37-38, italics in original).

The extent to which the Rasch model is applicable for a set of test data can be investigated by model-data fit procedures, and the results of these procedures can be used to evaluate model-data fit globally (over all the persons and items in a set of data) and individually (for each person or item). In the literature, there are several families of model-data fit procedures that are used to evaluate fit. Engelhard (2013b) mentions three: Pearson χ^2 , Power-Divergence, and Likelihood Ratio χ^2 . In this study, model-data fit was evaluated with methods based on the Pearson χ^2 , and specifically were Infit, Outfit, and Between (Bfit) Mean Square Error (MSE).

Infit, Outfit, and Bfit Mean Square Error

Generally speaking, Infit, Outfit, and Bfit MSE statistics provide information about the amount of variability in the response data compared to the variability that would be expected based on perfect fit to the Rasch model. Bond and Fox (2007), Engelhard (2013a), and Smith (2004) provide excellent details on the conceptual framework and procedures for calculating Infit and Outfit MSE. Smith (1985, 1986) provides the framework and procedures for calculating Bfit. The brief overview provided below is based on these references.

A response residual is the deviation of an observed response from what is expected or predicted based on the model. For dichotomous items, the responses are either 0 or 1, where 0 indicates that the response was incorrect and 1 indicates that the response was

correct. The expected responses are the (conditional) probabilities of a correct response derived from the model. In the case of the dichotomous Rasch model, the residuals can range from ~ -1 to $\sim +1$ (because the response probabilities can range from .00 to .99). To standardize the residuals, the residuals are divided by the standard deviation of the expected response probabilities for the item or person. These standardized residuals are used in the calculation of Outfit and Infit MSE.

Outfit MSE statistics provide the average standardized residual differences between observed and expected patterns in data. They can be calculated for either a person or an item. The person formulation of Outfit MSE (Engelhard, 2013a) is

$$\text{Outfit MSE}_n = \sum_i^L Z_{ni}^2 / L, \quad [2]$$

where

Z_{ni}^2 are the squared standardized residuals of person n and

L is the number of items.

In calculating Outfit MSE, the standardized residuals are squared before they are summed. Because Outfit MSE statistics are unweighted averages, they are sensitive to extreme unexpected residuals (outliers) (Smith, 2004).

Infit MSE statistics provide information-weighted, average standardized residual differences. They too can be calculated for either a person or an item. The simplified person formulation of Infit MSE (Engelhard, 2013b) is

$$\text{Infit MSE}_n = \sum_i^L Y_{ni}^2 / \sum_i^L Q_{ni} \quad [3]$$

where

Y_{ni}^2 are the squared residuals of person n ,

Q_{ni} is the variance of the expected response probabilities for person n on item i ,

$p_{ni}(1 - p_{ni})$, and

L is the number of items.

Because Infit MSE statistics are weighted averages, these statistics are less sensitive to outliers (Smith, 2004).

Bfit MSE tests the tenability of the Rasch model assumption that a person's achievement estimate for the total test should predict his or her observed scores on different subsets of items on the test. Bfit compares a person's expected scores on different subsets of test items with his or her sum total scores on the item subsets (Smith, 1986). A large Bfit value will be calculated if a person's achievement estimate from her performance on the total test cannot account for her performance on one or more of the item subsets. The item subsets for the Bfit statistic are established a priori and can be based on any grouping, such as the order of item presentation, item difficulty, or item content clusters.

The person formulation of the Bfit statistic is

$$\text{Bfit MSE}_n = \frac{1}{(J-1)} \sum_{j=1}^J \frac{\left(\sum_{i \in j}^{L_j} X_{ni} - \sum_{i \in j}^{L_j} E_{ni} \right)^2}{\sum_{i \in j}^{L_j} V_{ni}}. \quad [4]$$

In this formulation, J is the number of item subsets and L_j is the number of items in each subset (Smith, 1985). All other terms are the same as was defined for the Outfit and Infit statistics. The residuals of different item subsets are each summed, squared, and then standardized. Finally, these item subset values are combined to obtain one statistic per person.

Although Infit, Outfit, and Bfit MSE statistics are calculated for each person or item included in the Rasch analysis, they can also be averaged across a facet in order to provide easily interpretable indices of model-data fit. When data fit the Rasch model, the expected value of these statistics is 1.00 (Engelhard, 2013b; Smith 1986); but values can range from 0 to positive infinity. High MSE values indicate response patterns that are more haphazard than expected, and low values indicate response patterns that are more perfect than expected.

Infit, Outfit, and Bfit MSE statistics follow an approximate chi-square distribution (so they are not symmetric around the mean). The sampling distributions of these statistics, by which statistical inferences regarding fit can be supported, have also been shown to depend on sample size (Smith, 2004). The interpretation of the values of the person fit statistics, or in other words the values that indicate misfit, may be different for each analysis. In the previous literature, thresholds for defining misfit have been established using different methods. The most common are rule-of-thumb values, standardized versions of the MSE statistics, correction formulas (based on sample size), and simulation. The methods used to determine misfit for the applications of this dissertation are discussed with each application (i.e., in Chapters 3, 4, and 5).

Person Response Functions and Model-Data Fit

So far, it has been argued that person response functions can graphically illustrate the extent to which a person's observed response patterns resemble the expected response pattern from a measurement model. Person response functions may also represent a summary of the response pattern (for instance, the average function) for a particular group of students. In these cases, the shape and slope of the response functions provides

information regarding model-data fit. Further, it has been argued that PRF can provide information useful for diagnosing reasons for misfit or to provide clues as to which or how many items in a particular response pattern may elicit misfit. This information, when included along with the test score, can provide greater detail about validity.

In Figure 1, two person response functions were illustrated, an expected person response function and an observed person response function. It was explained that the expected person response function showed the pattern of responses that were expected by the IRT model and the observed person response function showed the pattern of responses that was present in the observed data. For this research, a visual comparison between the expected and observed PRF is conducted. At this time, I will introduce a different way to categorize PRFs that is based on how they are generated. This categorization is parametric PRF and non-parametric PRF.

Person response functions are comprised of a series of plotted values. How the values are generated provides a description of the type of person response function that is created. The PRF shown in Figure 1 are parametric PRF. They are relatively smooth and monotonic, and have no jagged hills or valleys. Compared to parametric PRF, non-parametric PRF are jagged, and many times include sharp hills and valleys. These differences in appearance are due to the way in which these plotted values are generated. The plotted values for the parametric functions are calculated using a model-based or empirically-derived mathematical formula. This mathematical formula dictates how the pattern of values will look, effectively forcing a shape onto the data (and sometimes ignoring the observed data completely). Conversely, the plotted values for the non-parametric functions are not calculated using an underlying logistic model. The jagged

appearance of non-parametric PRF occurs because the function is following the flow of the observed data.

The differences in how the parametric and non-parametric person response functions are created have implications for how they may be best used. Because a mathematical formula is imposed on the data, the parametric function shows a pattern that is expected by a given model (e.g., the underlying mathematical formula is obtained from the model) or a statistical summarization of the pattern that is observed in the data (e.g., a regression equation obtained from the data). However with parametric PRF, the fit of the PRF to the actual data must be explicitly examined because it is possible for the resulting PRF to not fit the observed data well. In other words, how accurately the PRF pattern reflects the pattern that is observed in the data must be evaluated using other sources, such as residuals.

By comparison, non-parametric PRF are more *free* to follow the observed data than parametric PRF because no underlying mathematical form dictates how the resulting non-parametric function should look. Consequently, non-parametric PRF show the pattern that is inherent in the observed data. The interpretation of non-parametric PRF is challenging because a frame of reference for judging the observed pattern is absent. In other words, without knowing what you are expecting to see, it is difficult to evaluate the pattern that you do see.

It is worth mentioning that both parametric and non-parametric PRF can be used to illustrate expected and observed person response functions. The choice of parametric or non-parametric PRF for an analysis depends on the context and purpose of that analysis. Some researchers of person fit use both parametric and non-parametric types of

PRF in their analyses (e.g., Trabin & Weiss, 1979; Engelhard, 2013a). In this research, I focus on discrepancies between the response pattern that is observed and the model-based, expected response pattern. Moreover, I use the Rasch model (a parametric IRT model) as the framework for describing how I expect the persons to respond to the items. For this reason, I use a parametric PRF to illustrate the model-expected response pattern. To illustrate the observed response patterns, I use both a parametric PRF (Application 1) and non-parametric PRF (Application 2 and 3).

In the following paragraphs, I discuss the details of the expected (parametric) person response function for the Rasch model. This is an important topic because in the three application studies, the Rasch-expected PRF will be used as the basis for comparison for the observed PRF. Because the type of observed person response functions are different across the application studies, the details of the observed PRF will be discussed more completely in the sections that include the application studies.

To explain the use of the PRF in the Rasch model, it is helpful to include a comparison between Rasch model expectations for person-item interactions, and the model expectations from the three-parameter logistic model (3PL, Birnbaum, 1968), another well-known IRT model. In the Rasch model, the only item level characteristic that is modeled is item difficulty, δ_i . In the 3PL model, two additional item parameters, item discrimination parameter (a) and a lower asymptote or pseudo-guessing parameter (c), are modeled. Item discrimination refers to the ability of the item to differentiate between individuals at different locations (achievement levels) along the latent variable. Pseudo-guessing refers to the probability that a person who is located at the low end of

the latent variable (low achievement) will obtain the correct response to the item by chance.

The 3PL for dichotomous items is written as

$$\phi_{ni1} = c_i + (1 - c_i) \frac{\exp[a_i(\theta_n - \delta_i)]}{1 + \exp[a_i(\theta_n - \delta_i)]} \quad [5]$$

where

a_i = discrimination parameter for item i , and

c_i = lower asymptote of the function (referred to as pseudo-guessing parameter for item i)

An important distinction for this study between the Rasch model (introduced in Equation 1) and 3PL model (Equation 5) is the implication of the model formulation on the item and person response functions. Panels A and B of Figure 3 show three item response functions for the Rasch and 3PL models. Panels C and D of Figure 3 show three person response functions for the Rasch and 3PL models. Represented on the y-axis for all of the functions in Figure 3 is the probability of giving the correct answer to a dichotomous item (i.e., $\Pr x=1$). For the item response functions, the x-axis represents the achievement levels of the persons. The increasing curve of the function shows that as the achievement levels of the persons (the locations) increases, the probability of giving the correct answer increases.

In the Rasch model, the item response functions do not cross because the discrimination values for the items must be approximately equal as laid out in the requirements of invariant measurement, and there is no overall guessing parameter for the item included in the model. In the 3PL model, the item response functions may cross

because the discrimination values for the items are not required to be equal and items can vary on their elicitation of pseudo-guessing.

For the person response functions in Panels C and D, the x-axis represents the difficulty levels of the items. The decreasing curve of the functions shows that as the difficulty of the items increases the probability of giving the correct answer decreases. The interpretation of the x-axis of the Rasch and 3PL person response functions follows a similar logic to the interpretation of the item response functions. The probabilities on the y-axis represent probabilities of giving a correct response for a single person or a group persons with the same locations (θ_n) as a function of different item locations (δ_i) (Carroll, Meade, & Johnson, 1991; Perkins & Engelhard, 2009).

In the Rasch model, the person response functions do not cross because the discrimination values for the persons, which represents how fast the probability decreases as items become more difficult, must be approximately equal as laid out in the requirements of invariant measurement and there is no overall guessing parameter for the person included in the model. In the 3PL model, the person response functions may cross because the person discrimination values for are not required to be equal and persons can vary on their propensity to guess at items for which they do not know the answer.

The parametric person response functions included in Figure 3 illustrate a pattern of responses that may be expected for each of the two IRT models. In real testing events, persons' observed response patterns are never as smooth and continuous as what is shown in the expected PRF. But, the extent to which a person's real response pattern conforms to the prescribed pattern evident in the expected person response functions is an

indication of person model-data fit, or person fit. (A similar idea pertains to model-data fit for items, but it is not discussed here.)

To compare the observed pattern to the expected pattern, two PRF must be drawn. Up until now, I have discussed the creation of the expected person response functions—the PRF derived based on a mathematical IRT model. Creating observed person response patterns follow a different general process and there are many ways to create them. Several broad categories are mentioned below. The specific methods for creating the observed person response functions are described in more detail later in the sections dedicated to the application studies.

A person's response to a set of dichotomously scored items will be a binary outcome, correct or incorrect. In educational testing situations, a correct response is coded as a "1" and an incorrect response is coded as a "0". Following the usual process of creating a PRF, one can envision that the items are ordered from easy to difficult and placed along the x-axis. When the actual correct responses and not the probability of giving a correct response are plotted along the y-axis, the resulting plot is comprised of some pattern of 1s and 0s because no other outcome is possible.

In an ideal testing scenario, which requires envisioning the underlying latent variable and items and persons placed on it, each person will give correct responses to all of the items that are located below (or easier than) their levels of achievement. The PRF for each of these persons would yield a step-pattern that consists of two horizontal lines (placed adjacent to each other and located at 1 and at 0) and a single vertical line that connects them. This step-pattern is equivalent to what is known in the measurement literature as a Guttman response pattern (Guttman, 1950). This response pattern would

exhibit good model-data fit and the vertical line represents the person's location on the latent variable.

In a more realistic testing scenario, the pattern of the PRF would yield a set of jagged peaks where persons respond incorrectly to some items that are too easy for them and respond correctly to some items that are too difficult for them. In reality, some model-data misfit is usually observed and even expected (Rasch 1960/1980). The psychometric issue surrounding validity is how much model-data misfit is too much to compromise score trustworthiness.

Making a judgment about the match of the jagged peaks of a raw response pattern and the expected person response function is difficult because the underlying pattern of responses is unclear due to the noisiness of the data. But, by smoothing the raw responses, converting proportion correct scores to a logit or normal deviate metric, or otherwise estimating the slope and intercept parameters of a response function by using the raw responses via a statistical framework, the underlying pattern of responses becomes clearer and the judgment about the match between the observed response pattern and expected response pattern becomes easier to make. This is the aim of creating observed person response functions. The three solutions mentioned above, smoothing, converting, or estimating, represent several broad categories of techniques used for creating and defining operational person response functions. The specific ways to operationally define and create an observed person response function are too plentiful to mention here. The specific methods for creating the observed person response functions in this study are described for each application study.

In summary, diagnosing person misfit to a measurement model and making a decision that too much misfit is present in a response pattern to warrant reporting a single test score require subjective and contextualized judgments. Visual comparison of expected and observed person response patterns (via a visual inspection of the person response functions) provides an alternative to the traditional statistical way of evaluating person misfit. The link between validity, model-data fit, and person response functions is explored theoretically by reviewing the relevant literature.

Uses of Person Response Functions

A summary of the methods and purposes of person response functions found in the educational and psychological literature is included in the following section. The summary includes a fairly exhaustive list of the researchers who have developed methods for constructing person response functions in chronological order. This list is organized into two categories: origins of person response functions prior to the development of IRT models and person response functions within IRT models. A table of these researchers and methods is included in Table 1.

Origins of Person Response Functions (Prior to IRT Models)

The bulk of research conducted about the methods and uses of person response functions occurred after the introduction of item response theory, yet the theoretical concept and first renderings of PRF date back to 1941 and classical test theory. Mosier (1940, 1941) wrote two articles which illustrated the theoretical links between test theory and psychophysics. In these articles, he describes the duality between a person's location and an item's location along a psychological continuum. He argues that a responses of a person to particular item (or stimulus situation in his words) is "not a function of the

individual alone, nor of the situation alone, but represents a *relation* between the individual and situation (1940, pp. 355-356, italics in original). To locate or find a person's level of achievement, one must first order the items from least difficult to most difficult. The achievement level of the person is defined as the level of item difficulty where success is 50%. That is, the location along the item difficulty continuum where a person gives a correct response 50% of the time is where his achievement level is found. (It should be noted that this framework for obtaining a person's test score was different from the traditional "number-correct" framework for obtaining a person's test score which was popular at this time.)

Because the purpose of Mosier's articles was to increase the academic dialogue between what he felt were two similar fields, test theory and psychophysics, the main thrust of the articles focuses on how the mathematical theorems of psychophysics (e.g., Thurstone's Law of Comparative Judgment (1927)) can be applied to educational and psychological tests with dichotomous responses by transposing the person-item response matrix. But, to demonstrate an alternate method to obtain a person's test score which takes into account the dualistic relationship between persons and items, Mosier (1941) uses a decreasing normal ogive with item difficulty included on the x-axis and the probability of a person giving a certain dichotomous response on the y-axis. The ogive shape of Mosier's PRF although very similar to the shape of IRT PRFs which come later, is used to denote the error in estimating the true score instead of a probabilistic relationship of giving a correct response between item and person locations along a latent variable. This PRF can be described as being motivated by classical test theory, instead of item response theory models, however, the attention paid to the idea of a *duality*

between the difficulty of the items and the achievement of the person as being the driving force behind a response pattern, and the mathematical basis for the shape and direction of the relationship, makes Mosier's PRF resemble an model-based (expected) person response function.

Carroll and Schohan (1953) independently introduced the idea of a person response function in the context of an end-of-course test for Navy officers. In what appears to be a technical manual, the authors provide details regarding the item development and selection of test items, the psychometric quality of the items, and the meaning of the test scores. Unlike Mosier (1941), the purpose of the PRF developed by Carroll and Schohan (1953), which they called operating characteristic curves, was to provide substantive curricular meaning to the test scores. Carroll and Schohan (1953) argue that by using the PRF approach, the Navy could obtain an estimate of the level of difficulty of tasks on which any particular Officer Candidate was likely to succeed versus the level of difficulty of tasks he was likely to fail. They argue this information is more meaningful to the Navy than information regarding a relative standing of a potential Navy Officer within a group of test-takers:

The scores on nearly all educational and psychological tests are usually interpreted in terms of relative, group standards...[these do] not specify the kinds of items the examinee can pass, nor the probabilities with which he will pass them. Suppose, therefore, one is interested in establishing a critical score such that students below that score will not be considered to have passed a naval officer candidate course. If one uses only relative standards, the decision becomes purely arbitrary. With the use of O. C. [operating characteristic] curves, however, one

can find that score which seems to stand at the critical point between acceptable and unacceptable candidates, in terms of what these candidates know or can do.

(Carroll & Schohan, 1953, p. 96, underline in original)

With the stated focus on curricular and criterion-referenced meaning, the authors argue not only for the use of PRF, but also appear to be promoting the importance of inferences about achievement that can extend beyond performance on a specific test.

Carroll and Schohan (1953) cited Lord's (1952) work using item characteristic curves and the general work of research of psychophysicists using varying intensity levels of stimuli to produce a response. This suggests that these authors were aware of measurement topics. Yet, their PRF does not appear to be inspired or based on this work. Carroll and Schohan (1953) do evoke the concept of an underlying construct of interest existing on a continuum, and that persons and test items differ with respect to how much of the construct they possess. Further they write that how much of the construct is possessed by the person and how much is induced by the item are both related to the probability of the person giving the correct response. These concepts are central tenets of item response theory. Further, Carroll and Schohan's (1953) PRF is model-based (i.e., the normal ogive). But because IRT was not commonly practiced at the time of Carroll and Schohan's (1953) work, the PRFs appear to be created for the primary purpose of conveying substantive information about a test score to educational stakeholders. In other words, conveying substantive meaning about what a person knows and can do – validity.

Brunk (1981) introduces Bayesian least squares techniques to estimate person and item response functions. Although, Brunk cites Lord and Novick (1968) as inspiration

for using his PRF with mental testing, no overt connection is made between the information gleaned from person response functions and validity of inferences from a particular test score. Thus, it seems that the underlying purpose of using PRF in this article was as an alternative way to estimate a person's achievement level (or item's difficulty level) using observed data.

Person Response Functions in Item Response Theory Models

Weiss (1973) and his colleagues (Trabin & Weiss, 1979; Vale & Weiss, 1975, pp. 32-33) are often cited as the first researchers to create the person response function based on item response theory models. The explicit objective of the studies conducted by these researchers was to create and advocate for methods of computerized adaptive testing, which they believed could improve the accuracy of ability estimation over traditionally administered paper and pencil tests. A fortunate byproduct of their exploration into adaptive testing was the notion that some individuals are not well measured by adaptive tests. This led to person response functions, or subject characteristic curves as they called them, which helped elucidate students for whom adaptive tests did not work well.

Although the main purpose of the PRF developed by this group of researchers was to examine and explore test score appropriateness for individual or groups of test-takers in an adaptive test framework, later use of PRF extended beyond that initial setting. For instance, in the article titled "Fit of Individuals to Item Characteristic Models," Trabin and Weiss (1979) write that a

single summary score, while more parsimonious than a description of a testee's entire response pattern, may not reveal the operation of other factors on test-taking behavior, such as guessing, anxiety, cultural bias, or lack of motivation.

Thus, total scores on a test do not indicate whether that test is inappropriate for a certain individual or group of individuals. (p. 6)

With this statement, Trabin and Weiss (1979) argue that a single test score is not enough information to convey a complete picture of a person's test-taking behavior. They continue that person response functions can assist researchers in singling out individuals for whom the test provides a poor measure.

In using PRF to compare the fit between person responses to an IRT model, Weiss and colleagues also recognized the need for a visual and statistical basis for interpreting the observed response probabilities and person response functions. In other words, the observed probabilities of giving the correct response and the resulting (observed) PRF needed to be compared to some ideal or true index of model-data fit. For the statistical comparison, they suggested using χ^2 analyses using the IRT *model-expected* probabilities as the expected values and the observed response probabilities for the expected probabilities. In their methodology, the χ^2 statistics were calculated over item difficulty clusters, not for each item. Similarly, for the visual interpretation of PRF, they suggested using the IRT model-expected PRF as the basis for comparing the observed person response function (Trabin & Weiss, 1979).

Weiss and his colleagues are often credited with introducing the idea of PRF, but other researchers have extended the idea. Lumsden (1977, 1978) echoes the sentiments of Weiss and colleagues in that his PRF is a measure of person response reliability (consistency) across a set of items. Also echoing Weiss and colleagues, Lumsden argues that a sufficient statistic (e.g. sum score) is not enough to tell the whole story about a person's achievement level. He takes this idea a step further by suggesting that persons

with the same total score, but different response patterns may have different instructional or occupational needs. Person response functions are one way to provide additional information with which evaluate student test performance.

Although similar themes are found between Lumsden and previous researchers who used PRF, Lumsden's philosophy of a person's achievement level differs from the philosophy of the previous researchers included here. Lumsden adopts the view that a person's achievement level is variable across the items in a test. In other words, that achievement level may change from one group of items to another in the same test and that the items on a test are "perfectly reliable" (Lumsden, 1978, p. 19). From this viewpoint, measurement error is also a characteristic of the person, not the item (Lumsden, 1980). In contrast, most researchers using PRF assume that a person's achievement level is fixed across a particular test event. From this viewpoint, what appears to be fluctuation in a person's achievement level is attributed to measurement error that is due to person state characteristics, such as fatigue or boredom, or item characteristics, such as unclear or poorly written items. This distinction of philosophy is important because it situates Lumsden's PRF as a useful tool for enhancing test score interpretation, but not necessarily for evaluating fit to an IRT model.

Strandmark and Linn (1987) explored a different theoretical approach to the evaluation of model-data fit. Instead of applying goodness-of-fit tests to ascertain adequate fit to an IRT model, they devised a generalized IRT model with two additional person parameters that attempt to estimate the effects that are likely to cause poor person fit to an IRT model. The first additional person parameter that is included models the effect of a person's variability across a set of test items. It is empirically related to the

person reliability philosophy of Lumsden (1977, 1978, 1980). The second additional person parameter that is included models the propensity to guess on items (relative to omitting the items) to which the answers are unknown.

With Strandmark and Linn's (1987) model, it is feasible to conduct model comparisons between the complex model and a more parsimonious (nested) model. This is a clever way to *statistically* evaluate person fit to a proposed IRT model. Because their model is complex, however, the person response functions must be interpreted separately for each combinations of values used for the three person parameters (achievement, reliability, propensity to guess). Given this fact, it is likely that the person response functions yielded from this model may not provide easily interpretable graphical representations person fit or information regarding test score appropriateness.

Almost 30 years after introducing the idea of an operation characteristic curve (which was conceptually equivalent to a person response function) for interpreting test scores in a criterion-referenced context, Carroll (1985, 1989, 1990) and his colleagues (Carroll, Meade, & Johnson, 1941) revisited PRF. The major focus of their newly conceptualized PRF was still on score interpretation. Yet along with this goal is the idea of building clarity of the measured construct (i.e., ability) through understanding its relation to the item characteristics (i.e., difficulty). Carroll (1985) argues that a critical part of clarifying the construct is by measuring it with unidimensional items. He explains that a way to check that the items are unidimensional is by plotting PRF:

If a test is essentially unidimensional, person characteristic curves will form a family of generally parallel curves and will be useful in interpreting patterns of

ability...An excessive number of inappropriate patterns might also reveal an underlying multidimensionality of the test. (Carroll, 1988, p. 249)

Furthermore, they introduce PRF theory and how it can be used to estimate item and person parameters of a test. Carroll (1990) and Carroll, Meade, and Johnson (1991) use PRF to estimate item and person parameters. In PRF theory, the same mathematical form as a traditional item response theory is used. But whereas the parameters of traditional models explain item characteristics, such as discrimination and guessing, Carroll (1990) and colleagues (Carroll et al., 1991) include them as person predictors:

The person characteristic function employs the same mathematical model as customarily used in item response theory [3PL]...the person characteristic function differs from the item response theory model only in that the probabilities yielded by the equation are to be studied for a single individual... as a function of different values of b , for different tasks. (Carroll, Meade, and Johnson, 1991, p. 110)

In terms of how the parameters of the person response function should be evaluated, these authors argue that the average slope of the person response functions and its variance can be used to inform test creators about the level and range of student achievement the test is measuring (Carroll, Meade, and Johnson, 1991). This information can serve as a validity “check” on the test items that make up the test as being appropriately targeted to the testing population and can yield information about how the construct is being measured in a particular context.

Reise (2000) developed a method for studying person fit that attends to model-data fit and seeks to explain potential sources of person misfit. In Reise’s method, the

responses to the items are analyzed by a series of hierarchical linear models, where no predictors are added first, then item and person parameters obtained by an IRT model that has been previously fit to the data are entered into the multilevel model to explain variance in the item responses. Residual variation in the item responses is taken as evidence for poor person fit to the measurement model. Other person characteristics can be entered into the model to help explain the residual variation.

Although Reise's approach focuses on statistical evidence for person misfit, person response functions play a major role in the approach. Reise (2000) defines person misfit solely as variation in person slopes of the PRF, or in other words, variation that is beyond the variation expected by the IRT model. The parameters of each person's PRF are estimated, evaluated, then if necessary, graphically illustrated. By combining detection and explanation of person misfit, Reise's approach seems like a promising way to describe and explain person misfit and inform trustworthiness of a test score. Reise's approach is used in the first application in Chapter 3.

Like Reise, Meijer and his colleagues, Sijstma and Meijer (2001), Emons et al., (2004, 2005) have a long history of studying person misfit, both statistically and graphically. One of the major contributions of their research has been on advancing the use of non-parametric PRF as a tool for evaluating trustworthiness of a test score. Unlike parametric PRF, non-parametric PRF are not prescribed by an underlying measurement model (i.e., IRT model). Moreover, a non-parametric PRF is based on observed item difficulties, instead of based on estimations of difficulty along an unobservable latent variable. Although a non-parametric PRF does not require an underlying form, these authors argue that to be most useful for evaluating person fit in a meaningful way, the

items should be invariantly ordered (possess the same order of difficulty for all persons), non-increasing in proportion correct, and unidimensional. This argument is similar to the argument made by Carroll, Meade, and Johnson (1991) and Carroll and Schohan (1953). It is noted that these psychometric properties that are advocated for describe those that are characteristic of the Rasch model.

Another idea developed and encouraged by Meijer and colleagues is the idea of a *comprehensive methodology* for person fit analysis (Emons et al., 2005). The authors argue that using a multistep approach to detecting and evaluating misfit provides more information to make a judgment about whether or not a score is misfitting (Emons et al., 2005). Their comprehensive methodology can be summarized as a three-step approach to detecting and evaluating person misfit. First, a general person fit statistic flags person's responses as misfitting or not. For the persons flagged as misfitting, a non-parametric PRF is computed. Then, for any places along the PRF that appear to be increasing (which is evidence of misfit) based on visual inspection, a second statistic is used to evaluate the misfitting response trend for statistical significance.

Moreover, Meijer and his colleagues have introduced several methods for creating person response functions. The PRF described in Sijstma and Meijer (2001) is based on the PRF of Trabin and Weiss (1983), but it is re-configured for non-parametric context (uses item proportion correct instead of an IRT item difficulty parameter for the discrete item subsets). The PRF described in Emons et al. (2004, 2005) uses the same method of PRF creation as is listed in Sijstma and Meijer (2001), but with a kernel smoothing step added to the process in order to make the pattern of responses to dichotomously scored

items more fluid (i.e., smooth). A second formulation of PRF described in Emons et al. (2004) is based on logistic regression with a kernel smoothing step.

The idea of using more than just a single value for examining misfit is a theme that resonates with the aim of this study. Using non-parametric PRF to evaluate person misfit is another similarity. However two major differences exist between this study and the studies presented by Meijer and his colleagues. First is that this study uses a parametric IRT model, the Rasch model, as the basis for the person fit and PRF comparison. In the United States, parametric IRT models are used in large-scale educational testing and the Rasch model is one of the most frequently used. This difference is noteworthy because the single expected response function for a person serves as a built-in frame of reference for the visual inspection of the observed PRF. The PRF created by Meijer and colleagues are not visually compared to model-based expected functions. Instead the parameters of the PRF are evaluated *statistically* for non-increasing values, which are indications of misfit because monotone decreasing values are expected (Emons et al., 2004, 2005). Secondly, the specific approaches used for computing the non-parametric PRF in the studies by Meijer and his colleagues and this study are different. For this study, methods for calculating the probabilities plotted in the PRF were chosen with the goal of being relatively easy to implement or relatively easy to conceptualize with minimal psychometric training.

Engelhard and colleagues have also presented ways of creating non-parametric person response functions and introduced ways of thinking about group-level PRF. Engelhard (2013b) used Hann smoothing to create non-parametric PRFs. He argues that a fit statistic alone does not provide enough information to make a good decision about

the trustworthiness of a test score. PRFs, specifically non-parametric PRF which have no underlying mathematical form, may provide authentic information regarding a student's response pattern and consequently misfit. The Hanning method is used to create PRF in applications 2 and 3 in Chapters 4 and 5.

Perkins, Quaynor and Engelhard (2011) extend person response functions to *subgroup* response functions in the context of international assessments. These group response functions can illustrate how subgroups of students may differ in their patterns of responses. They suggest that group response functions can assist in evaluating differential person response functioning (Johanson & Alsmadi, 2002) across groups of examinees, which could be indicative of model-data misfit for these groups. Walker and Engelhard (2014) suggested that person response surfaces, which are graphical representations of a person's responses across items that measure more than one dimension, can help inform score meaning. These surfaces may be particularly useful in conveying the trustworthiness of scores from game-based assessment contexts, where multidimensional information about student responses is collected.

Ferrando (2007, 2014) presents general approaches for investigating individual person fit in psychological and personality testing. First, Ferrando (2007) shows how individual person fit to congeneric test factor analytic model can be assessed using a procedure based on the *lz* and *lo* person fit statistics (Levine & Ruben, 1979). Then, he presents an approach based on Emons et al. (2005) to examining person reliability where the person's achievement level is considered to be variable during the testing event (Ferrando, 2014). Like the philosophy of Reise (2000), person misfit to a congeneric test factor analytic model, misfit connotes that the person's score is not interpretable in terms

of the construct of interest. Like the philosophies of Lumsden (1977) and Strandmark and Linn (1987), the reliability of a person's responses according to Ferrando (2014) is interpreted as parameter that can help explain his or her item response behavior. In both lines of research, Ferrando (2007, 2014) asserts that person response functions can provide complementary information about individual misfit. Moreover, his use of PRF is similar to that found in other IRT model-data fit studies where the function expected by the model is compared to the function that is observed from the data. Ferrando (2007) utilizes a two-step process to examining person fit, with statistical and graphical elements. Ferrando (2014) utilizes the comprehensive technique for examining person fit found in Emons et al. (2005) using global and local fit statistics and person response functions.

Summary

The ways in which person response functions have been created in the past and into the present differ, but the theories and uses of person response functions can be summarized into three themes. First, PRFs have been used to evaluate model-data fit at a *global* level. PRF can be used to inspect the match between a measurement model and the test taker population overall or between a measurement model and a specified group or groups of test takers (e.g., students with disabilities or different race/gender groups). The extent to which the expected PRF matches with the PRF observed for the group of test takers informs whether or not the essential properties of measurement invariance have been approximated well enough to move ahead with scoring and reporting.

Secondly, PRFs have been used to evaluate model-data fit at the *individual* level. PRF can be used to inspect the match between a measurement model and a single test

taker. The extent to which the expected PRF matches with the PRF observed for the person informs if the essential properties of measurement invariance have been approximated for that person well enough to trust that the score is a good representation of his or her achievement level.

A last theme that emerged was the use of PRF to *estimate person and item parameters* (Brunk, 1981; Carroll, 1990; Carroll, Meade, Johnson, 1987). For instance, in Carroll and colleagues' work, the average slope of the person response functions can be used to inform test creators about the level of achievement their tests are measuring and the variance of the average slope parameter can be used to inform the range of achievement levels the test is measuring. In Brunk's work, the posterior linear mean and variance of PRF are informed by both prior knowledge and real data. These extensions of PRF theory as an alternative technique for item and person parameter estimation did not appear to catch on in the measurement research community, but they were clever enough to be worth mentioning here.

Drawing together the common points from the past research regarding the utility of person response functions as a measurement tool, one could say that using PRF provides a way to contextualize score meaning with regards to the types of items that are included on a test. In this sense, PRF can help measurement professionals evaluate hypotheses regarding the nature of the construct being measured by the test. This supports the premise that person response functions can help promote the validation of inferences regarding the meaning of person scores.

This study continues and extends work in exploring person fit and person response functions as a way to validate test score inferences both at the global and

individual levels through model-data fit. In this study, I focus on the Rasch model and use a combination of real and simulated educational test data. The three applications that explore data fit to a Rasch model using person fit and person response functions are briefly summarized here:

Chapter 3

The first application applies a method for examining person fit based on multilevel logistic regression to a set of real large-scale educational test data. This method combines the benefits of statistical, graphical, and explanatory approaches to studying person fit and presents a way to explore global and individual person fit in large-scale testing programs. Further, the approach provides a way to construct person response functions that can help test practitioners visualize person misfit.

Chapter 4

The second application examines a two-step approach for examining person fit using person fit statistics and person response functions. Two person response functions were created, one based on non-parametric and exploratory data analysis and one based on the parametric expectations of the Rasch model. A visual comparison of the two PRF was conducted. A small dataset comprised of persons who used primarily guessing strategies to answer a 63-item multiple choice test was used to present person response functions that do not fit the Rasch model. Although these misfitting person responses incurred similar person fit statistics, their person response functions showed different response patterns. The study highlights the need for information beyond a single person fit statistic to understand student misfit and offer post-test advisement. The combination

of parametric and non-parametric person response functions that were used provide a way to help practitioners understand person misfit.

Chapter 5

The third application extends the two-step approach to examining person fit (with statistical and graphical procedures) to the computer adaptive test context. It uses simulated data, and a visual examination of person response functions of two groups of test-takers: Those test-takers whose responses fit the model and those test-takers whose responses did not fit the model. Because in computer adaptive tests (CAT) each test taker receives a different set of items, the traditional ways of examining model-data fit are limited. Person fit analyses are a way to examine model-data fit that is compatible with CAT because they evaluate how well each person's responses fit with the model. Person response functions provide a visual representation of misfit and can help researchers and practitioners understand misfit in these tests.

Chapter Three: Exploring Person Fit with an Approach Based on Multilevel Logistic Regression³

This chapter focuses on a promising method for detecting and conveying person fit for large-scale educational assessments. This method uses multilevel logistic regression (MLR) to model the slopes of the person response functions, which is a potential source of person misfit for IRT models (Reise, 2000). I apply the method to a representative sample of students who took the writing section of the SAT (N=19,341). The findings suggest that the MLR approach is useful for providing supplemental evidence of model-data fit in large-scale educational test settings. MLR can be useful for detecting general misfit at the global and individual levels, and for graphically conveying general misfit of individual students to educational stakeholders using person response functions. However, as with other model-data fit indices, the MLR approach is limited in providing information regarding only some types of person misfit.

Justifying the proper interpretations and uses of test scores is central for establishing test validity (AERA/APA/NCME, 2014). One way that validity evidence is established is by examining the psychometric properties of persons' responses to the test items. Current item response theory (IRT) methods for examining psychometric quality focus on how items perform on their own and how the items perform together as a set. Indices of item-level model-data fit help test creators make decisions about what items provide useful measurement of the construct for a group of test-takers.

³ This is an Accepted Manuscript of an article published in *Applied Measurement in Education*:

Walker, A. A., & Engelhard, G., Jr. (2015). Exploring person fit with an approach based on multilevel logistic regression. *Applied Measurement in Education*, 28:4, 274-291, doi: 10.1080/08957347.2015.1062767.

Standard computer programs that implement IRT methods (e.g., Winsteps, eRm) also yield indices of *person*-level model-data fit. These person fit indices describe the response consistency of persons across the items that make up a test, and they can also help test score users to pinpoint persons or groups of persons who may differ unexpectedly in the way they respond to items. Person fit indices are used in the quality control checks for large-scale assessment practice, but they are not used to the same extent as item-fit indices. One reason for this is because person misfit can occur in many different ways and for many different reasons. For instance, a student could be nervous and fumble over the first set of items he or she encounters and answer many easy items incorrectly. Or, a student may have special knowledge of a topic (e.g., medieval weaponry and warfare) which allows him or her to answer a set of difficult items correctly. The many ways in which person misfit may occur makes the detection and resolution of it challenging.

However, like item fit, person fit is an important aspect to examine for establishing the validity of test score inferences. Person misfit implies that a person's responses to the test items are not solely determined by his or her achievement level. This is a sign that the obtained test score, and consequently the construct interpretation (e.g., Proficient or Not Proficient), may be a poor representation of what he or she actually knows and can do and what he or she should learn next. In other words, person misfit indicates that the meaning of the student's test score is questionable.

When the meaning of a student's test score is questionable, the use of the test score to make educational decisions for that individual student is questionable also. For examining the validity of test score interpretation and use in educational achievement

testing, it seems reasonable that the evaluation of person fit at the individual level in addition to the global level indices of model-data fit should be conducted. Moreover, an indicator of how trustworthy a score is for a student given the purpose of the test should be included in test score reporting.

In this study, I focus on a promising method for detecting and conveying person fit for large-scale educational assessment practice. I apply this method to a representative sample of students who took the writing section of the SAT. More details about person fit and the approach used to detect and convey person fit are provided in the following sections.

Person Fit and Validity

With the development of item response theory, model-data fit both at the item- and person- levels became an important area of research because adequate model-data fit is essential for the invariant properties of IRT models to be achieved (Swaminathan, Hambleton, & Rogers, 2007). But studying person fit within an IRT context is important for more reasons than its association with overall model-data fit. As is evident in the classic measurement writings of Thurstone (1926), Mosier (1940, 1941), Cronbach (1946, 1950), and Guttman (1950), person fit is also important because it has implications for the meaning and appropriate use of test scores. Stated succinctly, person fit informs test score validity. Cronbach (1950) attests to this sentiment when he writes that eliminating extreme cases of person misfit “has the disadvantage of throwing out numerous subjects, but it is vastly better than treating the subjects as if the scores were valid” (p. 26).

In major testing programs today, global model-data fit is often achieved due to the advances in item development, item banking, and test construction that have been made over the past 50 years. But despite global fit to an IRT model, it is possible and probable that responses from some persons within the set may not fit with the model's expectations (Rudner, Bracey, & Skaggs, 1996). In such cases, acceptable levels of fit for most individuals would be observed, but poor fit for a few individuals would also be observed. For these individuals, the obtained test score may not be a good indication of what they know, can do, and what they should learn next.

Identifying persons whose item response patterns are not reasonable given their estimated achievement level and describing where and how scores deviate from the expected pattern remains an important step for establishing validity in current testing practice (Messick, 1995). Current person fit research reflects this idea. For example, Cui and Roberts (2013) used a person fit statistic and student verbal reports to validate a cognitive model for a diagnostic assessment. Petridou and Williams (2010) used person fit statistics and teacher verbal reports to inform the extent to which test scores were justified for the students taking a test. Walker and Engelhard (2014) suggested that person response functions, which are graphical representations of person fit, can help inform score inferences from game-based assessments. These and other researchers demonstrate how attending to person fit during all stages of test development is essential for establishing validity of test score inferences.

Statistical, Graphical, and Explanatory Approaches to Studying Person Fit

In the recent literature on person fit, three broad categories of research have emerged: a) developing and refining person fit statistics, b) creating ways to graphically

visualize person fit, and c) modeling or explaining person misfit using linear regression or hierarchical linear models. The category that has been examined the most is the first one listed above, the development and refinement of person fit statistics. Karabatsos (2003), Meijer and Sijtsma (2001), and Reise (1990) provide reviews of the many person fit statistics that are available for detecting and evaluating person fit.

Within an IRT context, person fit statistics generally quantify residuals, the deviation of each response given to an item from the response (probability) that is expected based on the underlying measurement model. These statistics are necessary for practical use because they can signal when misfit occurs, but they are somewhat limited in a formative or diagnostic capacity because they do not indicate where the incongruent responses occur and for what reasons. On the opposite side of the spectrum, residual analyses, which examine individual deviations of observed and expected response probabilities for each item, show where incongruent responses occur in great detail. Residual procedures may have limited practical applications because they provide too fine a level of granularity in flagging person misfit.

Graphical methods for displaying person fit offer benefits over purely statistical approaches in that they can highlight where incongruent response patterns are in relation to a person's entire response pattern (Emons, Sijtsma, & Meijer, 2005; Engelhard, 2013a; Perkins, Quaynor, & Engelhard, 2011; Sijtsma & Meijer, 2001; Trabin & Weiss, 1979). Explanatory approaches offer the potential benefit of modeling effects associated with the misfit that may lead researchers to understand why the misfit occurred (Lamprianou & Boyle, 2004; Lamprianou, 2010, 2013; Petridou & Williams, 2007; Woods, 2008; Woods, Oltmans & Turkheimer, 2008). Graphical and explanatory procedures can provide more

information regarding misfit than a person fit statistic alone, but not as much information as a residual analysis. In this sense, graphical and explanatory approaches may have the potential to provide the right amount of information to be practically feasible and diagnostically useful.

Combining Approaches Using Multilevel Logistic Regression

Reise (2000) outlined an approach for examining person fit that combines the benefits of the statistical, graphical, and explanatory methods. He conceptualizes person fit in an explanatory multilevel IRT framework (De Boeck & Wilson, 2004) where the item responses are considered to be repeated samples drawn from person achievement levels. Using this framework, Reise (2000) uses multilevel logistic regression (MLR) to examine the extent to which person responses to the items misfit an IRT model. The person slopes and intercepts of the person response functions are treated as random effects where un-modeled and systematic variability can be detected. The estimated slope parameters obtained from the MLR can be compared to the slope that is expected under the measurement model to gauge model-data fit. Person response functions can be graphed for each person using the empirical Bayes slope and intercept parameters obtained from the MLR, and with this graphical step, it is possible to see where misfit in the person's response pattern may be occurring. The approach also allows for any systematic variability across person slopes and intercepts to be modeled using explanatory predictor variables (Reise, 2000).

The benefits of using multilevel logistic regression include providing information regarding overall model-data fit, illustrating evidence for test score validity via a graphical format, and potentially providing diagnostic fit information in terms of where

students responses misfit the model and why. There have been only a few studies that have evaluated or applied this method to the study of person fit (Conijn, Emons, van Assen, & Sijtsma, 2011; Wang, Pan, & Bai, 2008; Woods, 2008; Woods et al., 2008). The results have been largely positive. None of the previous studies used real, large-scale education data to explore the approach, and it is conceivable that it may be feasible for use in large-scale educational assessment contexts.

Purpose

The purpose of the first application is to explore the use of multilevel logistic regression (MLR) for detecting and conveying person misfit within the context of a large-scale educational data set. The two research questions addressed by this study are:

1. Are persons responding to test items designed to assess writing achievement as expected based on the Rasch measurement model?
2. How can multilevel logistic regression analyses contribute to the detection and understanding of person fit in large-scale educational testing?

As is outlined in Reise (2000), the study follows a two-step procedure. First, the items and persons are calibrated using an IRT model -- the Rasch model. Second, MLR is used to assess person fit to the Rasch model by examining person slope variation beyond what is expected by the model. The Rasch model and the MLR approach used in this study are introduced next.

Rasch Model for Dichotomous Items and Person Response Functions

Rasch measurement theory represents observed item responses with two estimated parameters: item difficulty and person achievement. The dichotomous formulation of the model is

$$\phi_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad [6]$$

where ϕ_{ni1} is the conditional probability of person n with ability level θ giving the correct response to item i with difficulty level δ . As with other IRT models, the functional relationship between item difficulty and person achievement for the Rasch model is often described with item characteristic functions and person response functions. A person response function is analogous to an item characteristic function except that the x-axis represents the difficulty levels of the items that make up the test instead of the achievement continuum. The y-axis represents the probability of the student giving the correct response. The monotonic *decrease* of a person response function shows that as items become more difficult, the probability of a person giving a correct answer decreases. The location of the curve along the x-axis relates to the person location on the latent variable, and the slope of the curve represents how fast a person's probability of giving a correct response decreases as items become more difficult.

Under the Rasch model, the slopes of all person and item functions are assumed to be equal (Andrich, 1989; Bond & Fox, 2007; Engelhard, 2013b; Lumsden, 1980) and they are set to 1.00 in practical applications of the Rasch model. The person response pattern that fits the Rasch model well exhibits a steady decline (with a slope of 1.00) at the location where the probability of giving a correct response is 0.50 (the person's estimated achievement level). For a response pattern that does not fit the Rasch model, alternate shapes and slopes of the PRF may be observed. For instance, a flat response function may be indicative of misfit to the Rasch model, because it would indicate that as

items become more difficult, the probability of the person of giving a correct response remains approximately the same.

Multilevel Logistic Regression Model for Person Fit

Reise's (2000) MLR approach models the intercept and slope parameters for each person's response function using each person's scored responses to the items on the test. The intercept represents the person's achievement level and the slope represents a measure of a person's response consistency based on the expectations of the underlying measurement model. Reise (2000) defines person fit solely in terms of the slope of the estimated person response functions. MLR analysis is used to detect and explain the variation in these slopes. By modeling the scored item responses as being nested within persons, the hierarchical structure of the test data is taken into account as well as the dependence of the level-one units of measure (items) with the level-two units (persons) (Raudenbush & Bryk, 2002).

Method

Participants

A five-percent random sample of students was selected from the total population of test takers from the October 2009 administration of the SAT ($N \sim 400,000$). Engelhard, Wind, Koblin and Chajewski (2014) reported close correspondence between the sample and the total population with regards to demographic composition; thus supporting the inference that this random sample ($N=19,341$) is a good representation of the population of SAT test takers during the October 2009 administration. Although approximately 19,000 respondents may be considered to be small in some large-scale assessment programs, this sample size is similar to the large-samples used in many statewide

assessments. Additionally, the method used in this study to examine person fit can be scaled-up to include larger datasets.

Instrument

The SAT is designed to help colleges and universities identify students who could succeed at their institutions, and to connect students with educational opportunities beyond high school (College Board, 2011a). Data from the SAT writing section was examined in this study. It is comprised of 49 multiple-choice items and one essay prompt. The 49 multiple-choice items are categorized into three areas: improving sentences, identifying sentence errors, and improving paragraphs (College Board, 2011b). For this study, only the 49 multiple choice items of the SAT writing section were used.

Procedure

Each item was scored as “1” if answered correctly and scored as “0” if otherwise⁴. The item responses were examined for person-level model-data fit using a two-step procedure (Reise, 2000) which is described below. It is noted that item response theory (IRT) in general, and the Rasch model in particular were not used in the original development, analysis, and scaling of the SAT.

Step 1: Rasch model item calibration and person measurement. First, the test data were fit to the Rasch model using the Extended Rasch Modeling (eRm) package (Mair & Hatzinger, 2007a) for the R platform (R Development Core Team, 2006). The eRm package was chosen for the Rasch analysis because it uses conditional maximum likelihood estimation (CMLE) procedures. It has been argued that CMLE is conceptually close to Rasch’s (1980) idea of specific objectivity (Mair & Hatzinger, 2007a, 2007b),

⁴ This practice represents “rights-only” scoring. This scoring method was not used to operationally score students’ responses during the October 2009 SAT administration.

and unlike marginal maximum likelihood estimation, CMLE does not require specifying a density function for the distribution of the person parameters. This detail is useful when real data are being used, and the underlying distribution of person measures is unknown.

In addition to providing individual item and person residuals and fit statistics, the eRm package also includes Infit and Outfit Mean Square Error (*MSE*) statistics which are used to assess model-data fit (Mair & Hatzinger, 2007a). Infit and Outfit *MSE* statistics summarize residuals, and as such they provide information about the degree of deviation between persons' (or items') observed responses and the expected probabilities for a correct response based on perfect fit to the Rasch model. High *MSE* values indicate response patterns that are more varied (noisy) than expected and low *MSE* values indicate response patterns that are more muted (less varied) than expected.

Infit and Outfit *MSE* are both averages of the standardized residual variance, but are calculated in different ways. Outfit *MSE* is the average of the standardized residual differences between observed and expected probabilities. It tends to be sensitive to outliers or extreme unexpected response patterns. Infit *MSE* is a weighted average of the standardized residual differences using each person (or item) variance as the weighting constant. It tends to be less sensitive to outliers (Engelhard, 2013b). The values for these fit statistics range from 0 to infinity, with an expected value of 1. (Engelhard, 2013b, Smith, 1991).

Step 2: MLR for Person Fit. The ordinary logistic regression models the probability of success on a dichotomous item (Reise, 2000) as

$$P_{ij}(Y = 1 | \delta_{ij}) = \frac{\exp(\beta_{0j} + \beta_{1j}\delta_{ij})}{1 + \exp(\beta_{0j} + \beta_{1j}\delta_{ij})}. \quad [7]$$

To make the relationship between this formulation and the multilevel logistic regression model formulations more apparent, Equation 7 can be written as $\ln(P_{ij} / Q_{ij}) = \beta_{0j} + \beta_{1j}\delta_i$ where $Q_{ij} = 1 - P_{ij}$ (Reise, 2000). Using the item calibrations and person measures obtained in Step 1, three two-level logistic regression analyses were used to evaluate person fit to the Rasch model. It is noted that same set of data used to calibrate the items and measure the persons is used in the MLR step of the person fit analysis. The specific multilevel models used in this study are described below.

Multilevel Model I. The first multilevel model is a random coefficients model:

$$\ln(P_{ij} / Q_{ij}) = \beta_{0j} + \beta_{1j}\delta_i$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

where

$\ln(P_{ij} / Q_{ij}) = \log$ odds of item i being answered correctly by person j ,

β_{0j} = the expected log odds item i being answered correctly by person j when the difficulty of item i is 0,

β_{1j} = the expected change in log odds of item i being answered correctly by person j when item difficulty increases by one unit, and

δ_i = the difficulty of item i .

At level 1, the person's response to an item is a function of a person intercept and a person slope (the relationship between item difficulty and a correct response). Item difficulty is centered at zero from Step 1 (for CMLE); thus the intercept has a meaningful interpretation. At level 2, the person-level intercept and slope parameters are treated as outcome variables. The γ_{00} of the level 2 equation represents the grand mean of the level

1 intercept, which is the overall average log odds of answering an item correctly when item difficulty is 0. The grand mean slope is represented by γ_{10} in the level 2 equation and is the overall average change in log odds when item difficulty increases by one unit. Each person's deviation from the grand mean intercept and slope are represented by the u_{0j} and u_{1j} , respectively.

No second level predictors are included in Model I. This model captures the amount of variation in intercepts and slopes that is observed across person responses to the items and the extent to which the grand mean of the slopes (γ_{10}) is a good representation of the change in person responses as item difficulty increases (Reise, 2000). In other words, it answers the question of whether or not the relationship between item difficulty and giving the correct response differs across persons.

Multilevel Model II. The second multilevel model is an intercepts-and-slopes-as-outcomes model. The level one equation remains the same. For the intercept level two equation, the estimated achievement level for each person, $\hat{\theta}_j$, is added to the model. The slope equation remains the same as in Model I:

$$\ln(P_{ij} / Q_{ij}) = \beta_{0j} + \beta_{1j}\delta_i$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\theta_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

In Model II, γ_{00} still represents the grand mean intercept when item difficulty is 0. The term u_{0j} represents an individual person's residual variation from the grand mean intercept after controlling for his or her achievement level. The γ_{01} represents the relationship between achievement level and β_{0j} .

Model II allows the person intercept coefficients to be explained by achievement level and permits inspection of any remaining variance not explained by the person's achievement level. In the context of a Rasch model, significant residual variation in person intercepts after achievement is introduced into the model signifies that another, potentially irrelevant, construct is influencing person responses to the items (Reise, 2000). This could signal poor overall model-data fit.

Multilevel Model III. In the third multilevel model, the level one equation is the same as in Models I and II. At level 2, the u_{0j} term, which represents an individual person's residual variation from the grand mean intercept after controlling for achievement level, is removed. In this model, the intercepts can vary across individuals, but they are explained in full by the person's achievement level. The intercept equation is justified when the observed variation in the intercepts can be completely accounted for by a level-two predictor. Again, the second-level slope equation remains unchanged:

$$\ln(P_{ij} / Q_{ij}) = \beta_{0j} + \beta_{1j}\delta_i$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\theta_j$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

From Models II and III, empirical Bayes estimates of individual's intercept and slope coefficients can be obtained. Person response functions can be created and inspected for individual person fit using these estimated parameters (Reise, 2000).

The three multilevel logistic regressions in this study were conducted with HLM for Windows, software version 6.08 (Bryk, Raudenbush, & Congdon, 2009), and used penalized quasi-likelihood (PQL) estimation. PQL estimation has been shown to produce biased estimates of the regression coefficients (Raudenbush, Yang, & Yosef, 2000) and

consequently, other estimation procedures that use more accurate approximations to maximum likelihood are often preferred. However, in the HLM software, the maximum likelihood methods do not provide univariate hypothesis tests that the level-two variances are zero. With PQL estimation, this test is available. Because I felt the hypothesis test for significant variation across slopes was necessary for the investigation, I chose to use PQL as our estimation method. As a check that our results were not severely biased, I also performed the analyses using the adaptive Gaussian quadrature approximation to maximum likelihood and compared the two sets of results.

Results

Step 1: Rasch model item calibration and person measurement

The results from the Rasch analyses for the 49 SAT writing section items indicate that the distribution of person measures was approximately normal with a mean person achievement estimate of 0.821 and a standard deviation of 1.210. This finding was important because it meant that the multilevel model assumption of normality of predictors was tenable. The mean person Outfit (0.983) and Infit (0.985) *MSE* statistics were close to the expected value of 1.000. The standard deviation of the Infit *MSE* was 0.167 and the standard deviation of the Outfit *MSE* was 0.505. This reflected that there was more variation in person responses when items were not well-targeted for persons than when the items were well-targeted.

The mean item difficulty was 0.000 with a standard deviation of 1.361. The mean Outfit and Infit *MSE* item statistics were 0.983 and 0.983, respectively, and these are close to the expected value of 1.000. The standard deviation of the Outfit *MSE* for items

was 0.211 and the standard deviation of the Infit MSE for items was 0.084. In general, this reflected an expected amount of variation in item responses.

Step 2: MLR for Person Fit

The parameters were estimated using the item calibrations and person measures obtained from the Rasch analysis of the same data used in Step 1 as level 1 and level 2 predictors. Two sets of parameters were estimated using both the PQL and Gaussian approximation to maximum likelihood (AGQ) procedures. The difference between the estimated coefficients using PQL and AGQ were minimal with all discrepancies being at the hundredths decimal place or smaller. To avoid redundancy, only the results obtained from the PQL analysis are presented.

Multilevel Model I. The primary purpose of Model I is to explore the statistical significance of intercept and slope variation across persons. This variation is included in the variances of the error terms u_{0j} and u_{1j} , $\hat{\tau}_{00}$ and $\hat{\tau}_{11}$. The $\hat{\tau}_{00}$ represents the variation of the individual intercept estimates from the grand mean intercept. The $\hat{\tau}_{11}$ represents the variation of the individual slope estimates from the grand mean slope. The closer the taus are to 0, the less variation is observed. These variance components can be subjected to hypothesis tests to see if the variability is statistically different from 0. Estimates of the average intercept ($\hat{\gamma}_{00}$) and slope ($\hat{\gamma}_{10}$) as well as estimates of the reliability of the ordinary least squares estimates are also obtained from Model I and provide a complete picture of the model's findings. The results are listed here, and they are summarized in column 2 of Table 2.

The value of the grand mean intercept ($\hat{\gamma}_{00}$) indicates that on average, the log odds of giving the correct answer to an item with a logit difficulty of 0 was 0.818. The

value of the grand mean slope ($\hat{\gamma}_{10}$) indicates that, on average, the log odds of giving the correct answer changes by -0.991 as item difficulty increases by one logit. It is noted that the grand mean slope is close to the expected value of -1.000 based on the Rasch measurement model.

The error variance for intercepts ($\hat{\tau}_{00}=1.247$) is large and statistically significant. The 95% range of plausible values (Van den Noortgate & Paek, 2004) for the intercept ranged from -2.259 to 3.895, which denotes much variation across individuals. The reliability of the OLS estimates of the intercept is high ($\hat{\beta}_{0jREL} = 0.870$) and means that there is enough variation across person intercepts to systematically distinguish between persons. Taken together these results imply that systematic individual differences in levels of writing achievement are observed in these data.

Model I also showed statistically significant error variance for person's estimated slopes ($\hat{\tau}_{11}=0.030$). The reliability of the OLS estimates of the slope ($\hat{\beta}_{1jREL}=0.207$) is much lower than that for the intercept, but statistical significance indicates that persons are varying from the mean on their individual changes in log odds when item difficulty increases by one logit. The 95% range of plausible values for the individual slopes ($\hat{\beta}_{1j}$) was from -0.654 to -1.328. This range suggests the person response slopes do not vary in direction, but they appear to vary in steepness. Because the Rasch model expects constant person response slopes (slopes equal to 1.00), these findings suggest misfit to the Rasch model from a person fit perspective.

Multilevel Model II. The parameter estimates for Model II are included in column 3 of Table 2. Model II introduces a second-level predictor, achievement ($\hat{\theta}$), to

predict the intercept, β_{0j} . Because the Rasch model was used to calibrate the items and persons and because the MLR intercept is conceptually similar to the Rasch achievement level, it is expected that all of the reliable variation observed in the person intercepts from Model I would be accounted for by the predictor, if the data fit the Rasch model well.

The changes in the person intercept parameters confirm this expectation. In Model II, the reliability estimate becomes close to 0 ($\hat{\beta}_{0jREL}=0.006$). The grand mean and the residual error variance become non-significant ($\hat{\gamma}_{00}=0.005$, $\hat{\tau}_{00}=0.001$). The estimate for $\hat{\gamma}_{01}$ (1.010) indicates a strong statistically significant, positive relationship between $\hat{\beta}_{0j}$ and $\hat{\theta}$. Approximately 99% of the variation observed in intercepts is explained by using $\hat{\theta}$ as a predictor. Taken together, these findings provide evidence that $\hat{\theta}$ explains practically all of the observed variation in person intercepts.

The grand mean slope and the error variance across person slopes change only slightly in Model II. The reliability of the OLS slope estimates is reduced in Model II. The substantive interpretation of these coefficients does not change from Model I.

Multilevel Model III. Model III was conducted because practically all of the observed intercept variation in Model I was explained by the introduction of achievement level, ($\hat{\theta}$), in Model II. For Model III, the u_{0j} term at level two was removed, signifying that the intercepts were considered to be fixed but varying across individuals and accounted for by achievement level. This intercept equation is theoretically compatible with the assumptions of the Rasch model because under the Rasch unidimensionality assumption, no variation in person intercepts is expected beyond what is determined by a person's achievement level. The second-level slope equation remained the same as

Models I and II with u_{1j} representing an individual person's deviation from the grand mean slope ($\hat{\gamma}_{10}$). As implied earlier, this slope equation is not theoretically compatible with the assumptions of the Rasch model because the Rasch model expects no variation across person slopes.

The parameter estimates for Model III are included in column 4 of Table 2. No substantial changes in the estimated coefficients are observed moving from Model II to Model III. The estimate of the grand mean achievement remained non-statistically significant, which suggested that fixing the residual variation to zero had minimal effects on the mean estimate. The relationship between β_{0j} and θ , represented by $\hat{\gamma}_{01}$, was 1.012. The grand mean slope coefficient for Model III was -1.026 and the 95% confidence interval for the true grand mean slope was [-1.032, -1.020].

The slope variance in Model III ($\hat{\tau}_{11} = 0.009$) was reduced from the variance observed in Model II, but it remained statistically significant from zero, implying that person responses were varying from the overall mean slope. Using the parameters from Model III, the 95% range of plausible values for $\hat{\beta}_{1j}$ was from -0.838 to -1.214. The reliability of the OLS slope estimates was reduced further from $\hat{\beta}_{1jREL} = 0.104$ in Model II to $\hat{\beta}_{1jREL} = 0.086$ in Model III. Because no reliable slope variation existed after Model III was conducted and because no other predictors should theoretically be included in a Rasch model to explain person responses to the items, no other models were created.

Person Misfit in MLR and Rasch Analyses

The substantive interpretation of the slope ($\hat{\beta}_{1j}$) in multilevel logistic regression defined by Reise (2000) is one of person response consistency. In the context of model-

data fit, the slope represents the extent to which persons give mostly incorrect responses to items located above their model-estimated achievement level and give mostly correct responses to items located below their model-estimated achievement level. This interpretation is conceptually similar to the interpretation of the Rasch-based fit statistics where the differences between the expected item responses probabilities (based on model fit given a particular achievement estimate) and observed item responses are statistically quantified.

To illustrate these interpretive similarities, the correlations between the estimated person slopes from Model III and the Rasch person fit statistics Infit and Outfit *MSE* were calculated and plotted in Figure 4⁵. The correlation between the slope parameters ($\hat{\beta}_{1j}$) and the Rasch Infit statistics was $r(19,268) = 0.978$, $p < .001$. The correlation between the slope parameters ($\hat{\beta}_{1j}$) and the Rasch Outfit statistics was $r(19,266) = 0.726$, $p < .001$ ⁶. These positive relationships indicate that as the estimated slope of the persons' response function ($\hat{\beta}_{1j}$) becomes flatter, the response patterns become noisier. Both relationships are strong, but the very strong relationship between the slope parameters ($\hat{\beta}_{1j}$) and the Rasch Infit statistic suggests that person misfit is being captured in a similar way across these procedures.

Person Response Functions

⁵ Rasch Infit and Outfit statistics were not calculated for 73 students who answered either all or none of the items correctly. These students were also excluded from the correlation procedure.

⁶ Two students with Rasch Outfit *MSE* statistic values greater than 20 were removed from the correlation procedure.

The results from the multilevel logistic regression analysis provided some evidence that the writing data fit the Rasch model, but not all students' responses were well-predicted by it. To explore the slope variation detected by the MLR, I utilized the graphical aspect of the multilevel logistic approach and plotted three person response functions. Using the 95% range of plausible slope values, I chose three students who had identical achievement levels and intercept estimates ($\hat{\theta}=.055$ and $\hat{\beta}_{0j}=.058$), but different estimated slope parameters ($\hat{\beta}_{1j}$). Student A had an estimated slope value of -0.895. This value was located on the flatter end of the range of plausible slope values. Student B had an estimated slope value of -1.008, which was in the middle of the plausible range of values and very close to the grand mean slope, $\hat{\gamma}_{10}$. Student C had an estimated slope value of -1.100. This value was one of the steepest slope values observed in the data.

The left panels of Figure 5 shows the person response functions drawn for these three students. The function represents the response probabilities that were estimated by Model III for each student. The probability of giving a correct response is listed on the y-axis. The item calibrations (in logits) are listed on the x-axis. The right panels of Figure 5 show the residuals for each student. The residual value is calculated by the observed response (either 1 or 0) minus the expected probability of giving the correct response (from the level I MLR equation) for a person with the same intercept and slope values.

The three person response functions shown in Figure 5 vary only slightly in steepness, and the shape of all three is smooth and ogive. The residual analysis, on the other hand, shows different levels of model misfit across the three students. For example, Person C has fewer large deviations compared to Persons A and B, and the pattern of the

residuals suggests that Person C's response deviations are mostly located around the middle of the item difficulty continuum. From a visual inspection of the residual plots, one may expect the PRFs of Persons A, B, and C to look more different than they do. That is, the differences in person misfit do not appear to be visually presented by the observed (estimated) person response functions.

Discussion

Feasible and accurate ways of detecting and examining person misfit are essential for validity in the same sense that ways of detecting and examining item misfit are essential for validity. In this study, I focused on one method for evaluating person fit at the individual and global levels that included a promising way to depict person misfit in large-scale assessments.

Research question one asked, *are persons responding to the multiple-choice items designed to assess writing achievement as expected based on the Rasch measurement model?* The findings from the Rasch analysis suggest that students tended to respond to the writing items as expected by the Rasch model. The average fit statistics, Infit and Outfit *MSE*, for items showed good fit to the Rasch model. The average fit statistics for persons also showed good fit to the model, although the relatively large standard deviation of the Outfit *MSE* statistic for persons suggested that some student responses were more varied than expected based on the Rasch model. This result makes sense given that in large-scale assessment there are many more persons being tested than items being used for the testing.

The findings from the multilevel logistic regression analyses extend the results of the Rasch analysis. The MLR analysis indicated that person intercepts did not exhibit

significant variation after controlling for estimated achievement levels ($\hat{\theta}$), which suggests that the persons with the same level of writing achievement were responding to the items in a similar way and that θ was conceptually equivalent to β_{0j} . The estimates of the grand mean slope were similar across all three models and the negative direction of the slope estimates indicated that students tended to answer difficult items correctly less often than easy items. Variation in the person slope estimates was observed and suggests that at least some students had person response slopes that were different from the value of 1.00 required by the Rasch model. From a practical standpoint, the range of estimated person slopes (-0.838 to -1.214), may be considered small (Linacre, 2000), and the estimate of the OLS reliability for person slopes (0.086), may be considered low. In practice, the final decision regarding Rasch model-data fit would evaluate person fit information as well as item-fit information.

Research question two asked, *how can multilevel logistic regression analyses contribute to the detection and understanding of person fit in large-scale educational testing?* Generally speaking, these results suggest that the MLR approach is sensitive to deviations from Rasch model expectations regarding the slopes (i.e., slope equal to 1.00). The approach may be a useful additional step for evaluating Rasch model-data fit in large-scale educational testing contexts. The results revealed that the MLR slope parameters (β_{1j}), when paired with a Rasch measurement model, detect person misfit similar to the ways that the Rasch fit statistics, Infit and Outfit *MSE*, detect person misfit. This finding suggests that the MLR is sensitive to unexpected responses to items targeted near and distant from a person's estimated achievement level. However, the very strong

relationship between person Infit *MSE* and MLR slope suggests that MLR captures misfit to the Rasch model in a fashion most similar to that of Infit *MSE*.

An implication of this finding is that all forms of person misfit may not be captured by slope variation in the MLR approach alone. This general idea, that a single fit statistic cannot capture all types of misfitting response patterns across all testing contexts, has been argued by other person fit researchers (Meijer, 2003; Smith, 1986, 2004). It is important to consider the benefits of implementing an extra procedure that yields similar information to a routine procedure based on Infit and Outfit *MSE*. The duality of Item Response Theory models suggest that both elements of the functional relationship (e.g., items and persons) play an essential part in obtaining appropriate and adequate measures. In current test practice, much time and energy is expended to ensure that the items have good model-data fit and support the validity of the intended uses and interpretation of the scores. During this process, person fit is sometimes examined, although not to the same extent as item fit. The multilevel logistic procedure used here provides an additional way to approach model-data fit from a person fit perspective in large-scale assessment programs.

In addition to providing corroborating evidence for model-data fit from a person-fit perspective, the MLR approach allows for the estimation of a range of plausible slope values for the person response functions. This range can be used as an effect size for global person misfit in a given testing population and also as a way to evaluate an individual person's level of misfit (i.e., by comparing individual misfit to what is plausible for the group of examinees). No IRT model will fit empirical data perfectly, and the magnitude of person misfit will differ based on each group. Thus, the range of

plausible slope values may be helpful for determining the extent of model-data fit from the person-fit perspective for both the entire testing group as well as for individuals within it.

In terms of the person response functions that were created in this study, the findings here indicate that MLR may provide a useful way to convey a binary decision regarding *general* person misfit, or in other words whether or not student responses fit with model expectations. However, for conveying *specific* person misfit information, such as possible locations along the item difficulty continuum where a student response pattern deviates from the model expectations, the person response functions created from MLR may not be useful. In all three of the person response functions illustrated in this study, the shape of the function was smooth and ogive. They looked very similar to each other. But given that the residual analyses for these three persons looked different, it seems that the parametric PRF did not pick up on the different patterns of person misfit. This mismatch suggests that using the person response functions derived from estimates of multilevel logistic regression analysis may not provide adequate information to capture and convey different types of misfit in diagnostic ways.

Other researchers have defined the relationship between the intercept and slope of the person response function to be person specific instead of model derived (e.g., Conijn et al., 2011). It is possible that this alternate conceptualization of person response functions may be more useful for illustrating location-specific person misfit for educational tests than MLR. Another alternative is to use a non-parametric graphical approach for constructing person response functions. Non-parametric approaches do not retain the strict requirements of an underlying mathematical form, which is usually

monotonically decreasing. Some researchers have argued that non-parametric graphical approaches provide the flexibility necessary to successfully illustrate misfitting response patterns (Emons, Sijtsma, & Meijer, 2004, 2005; Engelhard, 2013a; Woods, 2008) or to examine within person multidimensionality (Carroll, 1993).

In conclusion, the multilevel logistic regression approach used in this study did not detect very much person misfit in this particular data set. The underlying message of the approach deserves further attention: Model-data fit from global and individual person-fit perspectives provide important aspects of test score validity and by combining statistical and graphical approaches researchers can examine person fit more comprehensively.

As for implications for testing practice, it is well known by measurement scholars that test scores that are earned with misfitting response patterns are not meaningful, and they should not be used to make educational decisions. Yet, this message is not communicated to educational stakeholders that use test scores to make educational decisions about students or teachers. This gap between research and practice regarding person fit is one that we as measurement experts should consider addressing, and continuing to develop approaches for examining person fit and its implications for the validity of test scores for intended purposes.

Chapter Four: Exploring Aberrant Responses Using Person Fit and Person Response Functions⁷

The first application illustrated a combined statistical and graphical approach to examining person fit at the global and individual levels. In the second application a two-stage, statistical and graphical approach is used. In this approach, person response functions are used to supplement the interpretation of person fit statistics. This chapter shows that person response functions can provide information about *absolute* person fit to a model, whereas fit statistics provide information about *relative* fit, given the other persons in the testing group. This person fit information can assist practitioners in using and interpreting individual student scores appropriately.

Not all test scores are equally trustworthy for inferring what a person knows and can do and what she or he should learn next. Person fit analyses provide a description of the trustworthiness of an obtained test score, and these analyses can help researchers to evaluate validity globally for a set of data as well as for the individuals within the set (Embretson & Reise, 2000; Smith, 1986). In an item response theory (IRT) framework, person fit analyses detect persons whose test responses are not consistent with the IRT model.

In the literature, person fit statistics have been the most common way that person fit has been examined. Much research has focused on the development and refinement of person fit statistics (Lamprianou, 2013; Meijer, 2003; Petridou & Williams, 2007).

⁷ This is an Accepted Manuscript of an article published in the *Journal of Applied Measurement*:

Walker, A. A., Engelhard, G. Jr., Royal, K. D., & Hedgpeth, M. W. (2016). Exploring aberrant responses using person fit and person response functions. *Journal of Applied Measurement*, 17(2).

Person fit statistics are important because they flag persons whose test score may not reflect what they know and can do. But, person fit statistics do not tell the whole story. For example, persons who respond very differently to test items can earn the same person fit statistic. Moreover, it is possible for a misfitting person fit statistic to be indicative of a person who used guessing to answer the items, of a person who was nervous at the beginning of a test, or of a low-performing person who copied answers from a high-performing neighbor.

Person response functions (PRF) are graphical representations of the relationship between a person's probability of giving the correct response to the items on a test and the difficulty of the items on a test (Trabin & Weiss, 1979). They represent an alternate way to examine person fit and because they provide a graphical representation of person fit, PRF can provide researchers and practitioners with details that are not available from person fit statistics alone, such as where misfit occurs.

Recommendations for students may differ based on the reasons for their individual misfit. Comprehensive person fit information should be made available to educational stakeholders to help them make decisions regarding student achievement and instruction. Person fit statistics together with person response functions are a promising way to convey this information (de Ayala, 2009; Ferrando, 2007).

Purpose

The purpose of this study is to examine misfitting response patterns with person fit statistics and person response functions. It is anticipated that using both techniques can provide a more nuanced portrayal of misfit. There are several broad categories of misfitting response patterns that have been examined in the literature on person fit (e.g.,

cheating, plodding, and carelessness). In this study, I focus on the category of guessing and use a dataset that includes guessed responses to test items to explore it. The research question addressed in the study is how person response functions provide supplementary information to practitioners about person misfit?

Guessing, Person Fit, and Person Response Functions

For the purposes of this study, guessing is conceptualized as a randomly varying phenomenon that depends on the person and the item (Andrich, Marais, & Humphry, 2012; Adams & Wright, 1994; Smith, 1993; Waller, 1976, 1989). This conceptualization situates the examination of guessing under the purview of person fit. According to Rogers (1999), guessing happens “whenever an examinee responds to an item with less than perfect confidence in the answer” (p. 235). She lists three varieties of guessing behavior that persons may use when a guessing strategy is employed: blind, cued, and informed. Blind guessing refers to the situation where a person guesses at random. Cued guessing refers to the situation where a person refers to stimuli in the item stem or in other places of the test to select a response. These stimuli could be unintentional hints or intentional misleads. Informed guessing refers to the situation where a person uses partial knowledge to select a response (Rogers, 1999).

Person fit researchers have shown that person fit statistics are able to detect blind guessing under simulated test conditions, especially when a statistic that is sensitive to guessing under the particular IRT model used is chosen (Karabatos, 2003; Meijer, 2003; Smith & Plackner, 2010; van Krimpen-Stoop & Meijer, 2000). The other types of guessing listed by Rogers (1999), cued and informed, have not been studied as extensively in the literature, perhaps because these are not thought of as strategies that are

sustained throughout a testing event. For instance, on a well-developed test, cued guessing may reflect poor item selection for a couple of items out of the entire set. Both informed and cued guessing implies that the test-taker is drawing on some content knowledge, even if it is not complete or correct. These guessing behaviors differ from blind guessing because in blind guessing no content knowledge is used to answer the items (Smith, 1993). These also differ from blind guessing because using a blind guessing strategy could more realistically be used throughout the duration of a test event, for example by an unmotivated or under-prepared test taker.

These conceptual differences between the different types of guessing strategies can have implications for interpreting test scores. For students who blindly guess on all of the items, it is clear that the obtained score is not a good representation of their knowledge and skills. For students who guess based on cueing or partial knowledge (informed), a decision regarding the validity of the obtained scores is less clear.

Making a decision about whether observed misfit is *too much* misfit for the score to be considered trustworthy requires judgment (Drasgow, Levine, & Zickar, 1996). Person fit statistics can detect many types of measurement disturbances, including blind guessing, but these statistics may not provide enough detail regarding the nuances of misfit that can help a practitioner make a decision regarding the trustworthiness of the score. Moreover, the distributional properties of many person fit statistics have been shown to be sample-dependent (Hambleton, Swaminathan, & Rogers, 1991; Reise, 1990), which means that the person fit threshold values change from one testing event to another. An implication of this is that gauging *absolute* misfit in a given testing event is not possible using a person fit statistic alone.

Person fit researchers have developed methods for constructing person response functions (Carroll, 1990; Emons, Sijtsma, & Meijer, 2004, 2005; Engelhard, 2013a; Lumsden, 1977; Reise, 2000; Sijtsma & Meijer, 2001; Strandmark & Linn, 1987; Trabin & Weiss, 1979; Weiss, 1973). The most direct approach to doing this is to use one person response function to illustrate a pattern of responses that is observed in the data (i.e., plotting the observed responses) and use another person response function to illustrate what is expected of the data under an IRT model (i.e., plotting the expected responses). The match between the expected and observed PRF depicts the level of absolute fit of the data to the IRT model (de Ayala, 2009; Trabin & Weiss, 1979).

Some of these researchers have illustrated how misfit due to guessing or other causes (Emons et al., 2004; 2005; Engelhard, 2013a; Trabin & Weiss, 1979) can be illustrated using person response functions. Researchers have also reported that person fit statistics and person response functions provide complementary information (de Ayala, 2009; Ferrando, 2007; Nering & Meijer, 1998), which suggests that both could be used to inform model-data fit and test score validity.

In practice, there are often individual persons whose responses do not fit the model, despite adequate fit of the data to the model overall. For these persons, the use of their test scores cannot be justified. In large-scale educational achievement testing practice, however, person fit analyses, either with person fit statistics or person response functions, are not typically conducted at the individual test taker level (Cui & Roberts, 2013)⁸. Consequently, person fit information is not provided to the practitioners that use test scores to make important educational decisions. The general approach that is used to

⁸ In practice, person fit and item fit procedures at the global (all test-taker) level are conducted to ensure adequate model-data fit across all test takers in the testing event.

examine person fit in this study involves a direct evaluation of individual model-data fit. It illustrates a final effort to identify students whose test scores are questionable representations of their achievement.

The general method that is used to examine person fit in this study involves a direct evaluation of individual model-data fit supplemented with self-report data regarding guessing strategies. In other words, the methods illustrated in this study represent a final effort to identify students whose test scores should not be used to define what they know and can do and what they should learn next.

Of course, in practice, it is impossible to know unequivocally whether or not a person's response pattern is misfitting or the type of misfit it exhibits because in most testing programs all that is known about persons are their responses to the test items and the difficulty of the test items. The data used in this study provide a unique perspective in this regard. Information about how participants answered each test item was collected in addition to their responses to test items. Moreover, the participants in the study did not have direct opportunity to learn the course content of the test items. Instead, they were expected to use various guessing strategies to answer the items (Royal and Hedgpeth, 2013). Despite not knowing the test content, participants were motivated because a focus on learning about the test and test process was fostered among them. Participants were highly motivated to try and outperform the medical students for whom the test was designed, even though it was expected that their response patterns would exhibit varying levels of person misfit.

The experimental design used in this study does not reflect typical or best testing practices that use well-targeted and appropriate items administered to appropriately

trained persons (AERA/APA/NCME, 2014). This study was designed to deliberately create a situation that generates guessing. But, the resulting data allow for an authentic exploration of person misfit due to guessing behaviors.

Theoretical Framework

For this study, the Rasch model (Rasch, 1960/1980) was chosen to represent the test data. Person response functions are assumed to be non-increasing functions of item difficulty. Under the Rasch model, the difficulty of the items is assumed to be invariant across persons. This stability of the item ordering allows for a clear evaluation of fit using person response functions. Moreover, in an anchored analysis, the item parameters are treated as known. Although misfit to the model can still manifest as item misfit, a more appropriate interpretation in these cases is that persons are behaving unexpectedly (not the items). The pairing of the Rasch model with PRF provided a conceptual fit between the technique used for examining fit and the IRT model chosen for the analysis.

Rasch Model for Dichotomous Items

The Rasch measurement model defines a person's response to a set of dichotomous items as a function of two estimated parameters: the difficulty of the item and the achievement level of the person. The formulation of the dichotomous Rasch model is

$$\phi_{ni1} = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad [8]$$

where ϕ_{ni1} is the conditional probability of person n with an achievement level θ giving the correct response to item i with difficulty level δ . The Rasch model assumes that as items become more difficult, the probability of giving a correct answer decreases, and

that a person with a higher level of achievement always has a higher probability of answering any item correctly than a person with a lower level of achievement (Bond & Fox, 2007; Engelhard, 2013b; Wright & Panchapakesan, 1969).

The functional relationship (between item difficulty and person achievement) is often illustrated with person response functions and item characteristic curves. These functions are described as parametric because they represent the *expectation* of the model and this expectation is imposed on the observed data. For person response functions, the item difficulty continuum is included on the x-axis and the probability of a particular person giving a correct response to the items is included on the y-axis. Rasch PRF may shift along the x-axis to denote persons who have less or more achievement, but they will retain the same shape because the function is derived by the mathematic (parametric) form.

One way of examining model-data fit is by visually inspecting how closely an observed set of data fit with the expectation of the model via person response functions (and item response curves) (de Ayala, 2009; Embretson & Reise, 2000; Trabin & Weiss, 1979). The idea is to superimpose an observed PRF on an expected PRF, so that discrepancies can be visualized and interpreted. In creating an observed PRF the intention is to capture the underlying pattern present in the observed responses. Researchers have suggested the use of non-parametric person response functions to depict observed responses (Engelhard, 2013a; Emons et al., 2005; Walker & Engelhard, 2015) because unlike parametric PRF, a mathematical formula is not imposed on the data. Consequently, the shape of a non-parametric person response function may be different for each person and may have dips and peaks, which reflects variation in dichotomous

responses in addition to the shift along the x-axis, which reflects variation in achievement level.

In this study, the non-parametric person response functions were created with an approach based on Hann smoothing (Tukey, 1977; called Hanning henceforth). The Hanning process of creating person response functions uses responses near each other to create an informative path through a person's whole set of responses. Velleman and Hoaglin (1981) provide a version of the Hanning sequence that was modified for use in this study. It is described as

$$s_i = (y_{i-1} + 2y_i + y_{i+1}) / 4 \quad [9]$$

where s_i replaces the dichotomous response, y_i , and refers to the "smoothed" person's response to items y_2 through y_{i-1} , and the items are ordered by difficulty.

Another way to examine the fit of data to the Rasch model is by using fit statistics, such as Outfit and Infit mean square error (MSE), and indices for the reliability of person and item separation. Rasch Outfit and Infit mean square error (MSE) can be calculated for either an individual person or an item and they can also be averaged across these facets. Because the focus of this study is on person fit, Person Infit and Outfit MSE will be discussed here⁹. Person Infit and Outfit MSE summarize the difference between the observed probability of the examinee giving the correct answer and the expected probability of the examinee giving the correct answer. Person Outfit MSE statistics provide the average standardized residual differences between observed and expected patterns in data. The person formulation of Outfit MSE (Engelhard, 2013b) is

⁹ A discussion of item Infit and Outfit MSE can be found in Smith (1985) and Wu and Adams (2013).

$$\text{Outfit MSE}_n = \sum_i^L Z_{ni}^2 / L, \quad [10]$$

where Z_{ni}^2 is the squared standardized residual of person n on item i , and L is the number of items. Person Infit MSE statistics provide information-weighted, average standardized residual differences (Engelhard, 2013b):

$$\text{Infit MSE}_n = \sum_i^L Y_{ni}^2 / \sum_i^L Q_{ni} \quad [11]$$

where Y_{ni}^2 is the squared residual of person n on item i , Q_{ni} is the variance of the expected response probabilities for person n on item i , $p_{ni}(1 - p_{ni})$, and L is the number of items.

The values can range from 0 to positive infinity, but when data fit the Rasch model, the expected value of Infit and Outfit MSE statistics is 1.00 (Engelhard, 2013b; Smith, 2004). High MSE values indicate response patterns that are more haphazard than expected by the model, and low values indicate response patterns that are more perfect than expected by the model. Person reliability of separation for the Rasch model is equivalent to traditional measures of internal consistency, such as Cronbach's alpha or KR-20. It provides an indication of how well the test takers can be distinguished (or separated out) along the achievement continuum by the items on the test. High values indicate that test takers can be reliably distinguished from each other (by their levels of achievement) by the items. Low values indicate a small range of person achievement levels, such that test takers cannot be reliably distinguished from each other by the items. Item reliability of separation for the Rasch model provides an indication of the range of item calibrations that is obtained by the current sample of persons, or in the case of an anchored analysis, like the one in this study, the range of item calibrations obtained by

the original sample. It ranges from 0 to 1, where low values indicate a narrow range of item difficulties along the latent variable.

Method

This application analyzed data that was previously used to investigate the psychometric functioning of items from a typical first-second year medical school test at the University of North Carolina at Chapel Hill (UNC-CH) (Royal & Hedgpeth, 2013). Details regarding the participants, instruments, and procedures of the original study are included here.

Participants

Thirty-one professional staff persons from the Office of Medical Education of UNC-CH School of Medicine volunteered to participate in the study. The age of the participants ranged from 25 to 64 years with a mean of 41 (SD=11.86) and the number of years spent working in medical education ranged from 1 year to 28 years with an average of 8 years (SD=7.71). To be eligible for inclusion, participants had to hold at least a bachelor's degree and have no formal education or experiential training in the physical, life, health, or biomedical fields. These criteria for inclusion were necessary to obtain guessing behaviors on the test. Prior to participation in the study, the participants were informed that the test items were designed for first and second year medical students, and that during the study they would be asked about curricular content for which they were not prepared.¹⁰

¹⁰ Administering medical school test items to persons who have not had medical training could elicit frustration and disheartening. In this study, however, a focus on learning about the medical school testing process was fostered among participants, and IRB protocols were strictly followed with the participants. They were highly motivated to do their best.

Instrumentation

Medical Test Items Used in this Study

At UNC-CH, a web-based system called the **Medical Student Testing And Reporting System (MedSTARS)** was developed for creating and administering assessments that closely resemble the National Board of Medical Examiners (NBME) tests. All of the items in the MedSTARS item bank are created by medical school faculty for the purpose of testing the knowledge and skills of their students. Faculty are provided training opportunities for constructing high quality test items.

A 63-item test was constructed by selecting a mix of easy, moderate, and difficult items from a pool of test items designed for students in their first and second year of the undergraduate medical school program (N items=821). The 63 items were selected to vary by curricular content, faculty author, and difficulty.

Measuring Self-Reported Guessing Behaviors

The participants indicated the strategy used to answer each medical test item that was presented to them: *Please identify the strategy you used to answer the previous question from the options below:*

- 1) *I did not guess.*
- 2) *Informed guessing: I selected a particular answer based upon previous partial knowledge of the subject, or I was able to eliminate particular answer options based upon previous partial knowledge of the subject.*
- 3) *Cued guessing: I selected an answer based upon some sort of stimulus within the test such as wording cues, cues associated with item stems, choices among answer options, test-wiseness, etc.*

4) *Random guessing: I selected a particular answer by blindly choosing an answer.*

This question appeared after each item and provided self-report evidence from each participant about their guessing behaviors.

Procedure and Analysis

To mimic the same test conditions as students in the medical school, the items were administered to the 31 participants via the same electronic format and followed similar test administration procedures.

The test data were analyzed with the dichotomous Rasch model using the Winsteps computer program, version 3.72.3 (Linacre, 2011). Because the participant responses did not represent a target population of test takers, the item parameters used for the Rasch analysis were anchored to the difficulty calibrations established when the items were administered to the target population. To obtain these anchor values, the average proportion correct values for each item were converted to a logit scale by using the formula

$$\delta_i = \ln \frac{p}{1-p} \quad [12]$$

where p was the average p-value for the item from the item bank. In a subsequent step, these logit values were centered with a mean of zero and standard deviation set to 1.00.

Person fit to the Rasch model for the 31 participants was examined with Infit and Outfit MSE. Because the sampling distributions of Infit and Outfit MSE have been shown to depend on sample size (Smith, 2004; Smith, Schumacker, & Bush, 1998) an adjustment was used to construct a range of acceptable fit values that was appropriate for the size of the sample: $1 \pm 2\sqrt{2/N}$ (Wu & Adams, 2013). The calculation provided a

95% confidence interval of the Infit and Outfit MSE values. Values outside of this range indicated person misfit.

Next, two types of person response functions were plotted and superimposed on each other to facilitate the visual comparison between the observed response pattern and model-expected response pattern. The model-expected PRF was based on the Rasch model (Equation 8) and reflected a parametric approach. The observed PRF was based on Hanning (Equation 9) and reflected a non-parametric, smoothing approach. For the smoothing of the endpoints (i.e., items 1 and 63), an adaptation of Equation 9 was implemented: the endpoint was weighted by two, the closest point was weighted by 1, and the sum was divided by three. I repeated the Hanning procedure by one iteration for each raw score point earned (total raw score) as was done by Engelhard (2013a). This rule seemed to provide an adequate smooth while preserving the original pattern of the responses.

Results

Anchored Item Parameters

The original p-values for the 63 items ranged from 0.13 to 0.99 and the distribution was mostly uniform ($M=0.66$, $SD=0.23$). When converted to the logit scale and centered, the difficulty values ranged from -2.40 to 1.92, and the distribution was approximately normal ($M=0.00$, $SD=0.99$). For the subsequent Rasch analysis, the item calibrations were fixed to these values. Only the person measures were estimated from the Rasch analysis.

Scored Medical Test Item Responses

For the 31 participants, the total raw scores ranged from 13 to 33. The distribution was slightly positively skewed with an average number correct of 20.97 (SD=5.47). On the logit scale, person measures ranged from -1.60 to 0.15 with an average of -0.84 logits (SD=0.47).

Table 3 summarizes the findings of the Rasch analysis and Figure 6 includes the variable map of the item and person locations along the latent variable. The horizontal line on the variable map represents the raw score that would be expected for chance performance on this 63-item test. Although the spread and targeting of the persons and item calibrations look satisfactory on the variable map, the fit indices reveal poor model-data fit. The observed person responses were not consistent with the expected responses of the Rasch model (with anchored item parameters).

The person reliability of separation value was 0.49, which indicated that the persons could be distinguished from each other by the items only marginally. Because, this index is conceptually equivalent to Coefficient alpha, this value indicated that a lower level of internal consistency was observed *for these items with this testing group* than 0.70, which was the criterion used to select the set of items. The item reliability of separation was 0.66. Because the analysis was anchored using previously obtained item calibrations, this value indicated that (historically speaking) the set of items represented a moderately wide range of item difficulties to measure the achievement.

Large mean Infit and Outfit MSE values for the items and persons were observed. The average person Infit MSE was 1.22 (SD=0.12) and the average item Infit MSE was 1.33 (SD=0.82). The average Outfit MSE was 1.37 for both items and persons (SD=0.21 for persons and SD=0.83 for items). Both sets of these mean values were larger than the

expected mean values of 1.00, and indicated that the variation in participants' responses to the items was greater than what was expected by the model. Poor model-data fit was not surprising given that the participants were not representative of the target population for whom the items were designed. It was expected that the model and difficulty parameters of the items would not predict the responses to the items given by these test takers.

The individual person Infit MSE statistics ranged from 0.97 to 1.44 and the individual Outfit MSE statistics ranged from 1.00 to 1.95 (N=31). The larger range of Outfit MSE values compared to Infit MSE values has been noted by other researchers (for example, Smith, Schumacker, and Bush, 1998), but because of the small sample size and the poor item/person targeting, the fit values were difficult to interpret. The 95% confidence interval calculation from Wu and Adams (2013) produced a range of acceptable fit values from 0.49 to 1.51 for Outfit and Infit.

None of the 31 participants earned individual Infit values that exceeded the upper threshold of 1.51. Seven participants had Outfit values that exceeded 1.51. Because the overall fit of the persons to the model was poor, it can be inferred that these seven participants exhibited substantially worse fit than the average level of poor fit. The response patterns of these seven persons and for three others were visually examined with person response functions. The results of the visual analysis are provided later.

Self-Reported Guessing Behaviors

After answering each medical examination item, the participants were asked to indicate which type of response strategy they used to answer each of the 63 items on the test. With 31 persons answering 63 items, there were 1953 total item/person encounters.

Out of this total, there were 1948 valid response strategies reported (99.7%), the remaining five (0.3%) were missing. Only 30 of these 1948 were described by the participants as not guessing. Further, no individual participant indicated that he or she did not use guessing on more than 8 of the 63 items. This information provided evidence that the participants primarily used guessing strategies to answer the test items.

All participants used more than one type of response strategy to respond to the 63 items. With respect to the type of guessing that was reported, blind guessing was the most frequent (53.5%). Cued guessing was the second most frequent (30.3%) and informed guessing was the least frequent (14.6%). This same general pattern of reported guessing strategies existed within each of the easy, moderate, and difficult item clusters, although there was less of a gap between the number of reported blind and cued guesses (blind=49.1%; cued=34.2%) for the easy items than was reported for the moderate (blind=55.4%; cued=30.5%) and difficult (blind=56.0%; cued=25.3%) items. The constancy of the reported strategies suggests that the participants did not change their guessing strategy depending on item difficulty. The most likely reason for this is because the participants were unable to distinguish between items of varying levels of difficulty. When test content is unknown, every item seems difficult.

Person Response Functions

The idea that the same test score may not be equally trustworthy across persons who earn that score is the essence of person fit. I reasoned that selecting person response patterns that yielded the same achievement estimate, but that were obtained from different reported strategies would illustrate this idea in a clear way using person response functions. In this data set, three persons met this criterion. The PRF for three

persons with $\hat{\theta} = -0.72$, their standard error of measure (SEM), their self-reported guessing behaviors, and their Rasch person Infit and Outfit MSE statistics are shown in Figure 7.

On the PRF in Figure 2, the x-axes represent the anchored difficulty levels of the items on the test. The y-axes represent the probability of the person giving the correct response. The solid lines represent the smoothed (Hanning, non-parametric) PRF. The dashed lines represent the expected (Rasch, parametric) PRF for the achievement estimate of -0.72 . The asterisks represent the dichotomous responses to the items, where 0 represents an incorrect response and 1 represents a correct response.

The shape of the smoothed person response functions for the three persons are distinct. Person 23 reported using cued guessing on over half of the items, which means that he or she used unintended hints (or intended misleads) to answer more than half of the items. This person response function is better defined (e.g. monotonic), and according to the fit statistics, it fits the expected Rasch function better than the person response functions for Persons 29 and 7.

Person 29 reported using mostly blind guessing to answer the items. The person response function is flatter than what is expected by the model. This indicates that there is less change in the probability of giving the correct response as items change in difficulty than the Rasch model predicts. Person 7 reported using informed and cued guessing to answer the items, which means that partial knowledge and unintentional hints (or intentional misleads) were used to answer practically all of the items. The shape of this function is somewhat erratic and not well-defined. For instance, it starts off flat, but then around item difficulty -1 logit, jagged peaks and valleys are observed.

The PRF capture the nuances of person fit in a response pattern. Yet, it is noted that the responses given by these three persons represent moderate fit to the model because their Infit and Outfit values did not exceed the 95% confidence interval of fit for a sample of this size. To examine misfitting response patterns, person response functions for seven participants who earned extreme person fit statistics (i.e., exceeded the 95% confidence interval range) were created. These seven person response functions are included in Figures 8 and 9.

All seven of the misfitting persons reported using mostly blind guessing to answer the items on the medical test. Despite this similarity, the shapes of the observed PRF for these seven are very different. The only tangible similarity that can be noted is that all seven observed PRF diverge greatly from the expected Rasch function. Moreover, when compared to the PRF of the three moderately fitting persons in Figure 7, the PRF in Figures 8 and 9 look extremely erratic.

Discussion

The purpose of this study was to explore the use of person response functions for providing additional person fit information about real and known aberrant person response patterns. An experimental design was used to provoke guessing. Person fit statistics and PRF were computed and created to visually examine response patterns.

The findings of the Rasch analysis indicated that for these persons, their responses could not be adequately predicted by the estimated person measures and the anchored item calibrations. These findings of inadequate model-data fit highlight the importance of appropriate item targeting in educational achievement testing. The items included in this study were developed for a medical student population and the participants in the

study represented a different population. The high proportion of aberrant item response patterns observed in this study is due to the misalignment of the items to the tested population.

Yet in real operational settings, some test takers will provide responses that misfit the model even if many or most of the test takers provide responses that fit the model. The person response functions in this study provide examples of what misfit due to poor opportunity-to-learn or item mis-targeting may look like in an operational test setting. The PRF shown here illustrate the idea that person misfit is specific to an individual and highlight the role that person fit analyses play in examining model-data fit at the individual level.

The research question asked how person response functions provide supplemental information about misfit to practitioners. These preliminary findings suggest that observed (smoothed) and expected person response functions used in conjunction with person fit statistics capture different and complementary information about person fit that can be used to inform the inferences of test scores. Person fit statistics provide an overall indication of a person's fit to the model *relative to* the amount of misfit present in the testing group. The PRF provided a visual representation of *absolute* person fit to the model. The peaks and valleys of the non-parametric PRF (Hanning) captured the nuances of misfit. The parametric PRF (Rasch) provided the context for interpreting these nuances with respect to model fit. These findings support the findings of other researchers, that person misfit may not be easily detected by one person fit technique (Emons et al., 2005), and that person fit statistics and PRF can complement each other in

providing person fit information (de Ayala, 2009; Ferrando, 2007; Nering & Meijer, 1998).

It is plausible that different recommendations may best serve students with different response patterns or who have different reasons why their response patterns may be aberrant. Both absolute and relative person fit information can be useful for conveying the trustworthiness of a person's test score to practitioners. For the seven persons with misfitting responses, re-testing or treating the score as if something went awry during the testing event may be recommended. The erratic response patterns shown in Figures 8 and 9 suggest that the inferences based on the scores for these persons will not be trustworthy descriptions of what they know and can do. The mismatch between the model expectation and the observed pattern can be easily seen with the plotted PRF.

Person response functions can also assist practitioners in better understanding the unexpected responses in a particular misfitting pattern by identifying which items triggered the unexpected responses. For example, in a well-targeted test given to students who had adequate opportunity-to-learn, the observed PRF may highlight content that the student may know (e.g., the highest peaks of PRF 2 and 16 in Figure 8) or the content on which they struggle (e.g., the lowest point of PRF 10 in Figure 8). Other researchers have suggested using PRF for hypothesizing about the testing processes of students with unexpected response patterns (for example, Emons et al., 2004; Lumsden, 1977).

Conclusion

In practice, the extents to which students use content knowledge or other response strategies, such as guessing or cheating to answer operational test items is unknown. In well-developed, large-scale educational tests, it is expected that most person responses

will fit the model adequately, while only a few will not. Person fit analyses provide researchers and practitioners with one more piece of evidence that supports the interpretation and use of the resulting test scores for making educational decisions about individual students. Person fit techniques like those explored here may not yield as accurate information regarding validity as is yielded by truly qualitative inquiry methods, but unlike qualitative methods, the techniques explored here may provide a way to convey person fit information for a larger number of test takers. They can convey a missing piece of model-data fit information that is needed in large-scale educational assessment practice.

Chapter Five: Using Person Fit Statistics and Person Response Functions to Validate Theta Estimates from Computer Adaptive Tests¹¹

Application Three uses the same two-step, graphical and statistical procedure that was used in Application Two to explore person fit in a computer adaptive test context. In computer adaptive tests (CAT), person fit analyses are important to conduct because the traditional (paper-pencil) methods for examining model-data fit are limited due to test takers receiving different sets of items. In this application, five thousand person response vectors from a computer adaptive test were simulated under the Rasch model and inspected for person fit. Because the adaptive test data used for this application were simulated, the results from the analysis are preliminary. But the patterns of misfit illustrated in the application may help guide practitioners in making judgements regarding person fit, and consequently about the appropriate use of scores from CAT.

The practice of treating all test scores as if they are equally trustworthy could be problematic. Test scores from response vectors that do not exhibit adequate model-data fit may not be accurate representations of a student's knowledge, skills, or abilities. The use of global model-data fit procedures, such as item and person-fit analyses conducted using the set of test data can help ensure that test scores provide trustworthy measures of student achievement. Applying individual person fit procedures could further inform one's judgment regarding the validity of the inferences made based on a test score for an individual student.

¹¹ A derivative of this work is published in *Pacific Rim Objective Measurement Symposium (PROMS) 2015 Conference Proceedings*, Springer Science+Business Media Singapore Private Limited.

Currently, person fit procedures are not widely used in educational testing practice (Cui & Roberts, 2013). In paper-pencil tests, the context where the majority of educational testing occurs, global and individual *item*-fit procedures are conducted to ensure adequate model-data fit. For instance the stability of the banked item parameters is evaluated for a large group of test takers prior to test score reporting to ensure no major discrepancies between the prior and current calibrations are observed (Hambleton & Han, 2005; Hambleton, Swaminathan, & Rogers, 1991). Global person fit procedures are sometimes conducted during the test development and checking phases of paper-pencil tests, as additional quality check on the item calibrations (Embretson & Reise, 2000; Smith & Placker, 2010) or to augment differential item functioning analyses (Reise & Flannery, 1996). It is well-known that misfitting person responses can influence the quality item calibrations in item response theory (IRT) models.

For computer adaptive tests (CAT), these traditional quality-checking procedures have restricted utility because there is a limited amount of data available for each item with which to conduct them. In CAT, item parameters are considered to be known (van der Linden & Pashley, 2010), and they are derived by the particular item response theory model chosen to represent the data. But, each test taker can potentially receive a different set of items with which to measure her achievement. Each item is selected and administered a different number of times, and for a given assessment, may not be administered at all.

Person fit analysis is suitable for evaluating model-data fit in the context of CAT because it provides a quantification of how well the person's responses accord with the model used to calibrate the items and generate the achievement estimate. It is possible

for person fit analyses to be conducted for each examinee. In this respect, person fit could inform the judgment about whether or not the inferences made based on adaptive test scores are supported by evidence and theory (APA/AERA/NCME, 2014). With increases in the numbers of computer adaptive tests being administered (Chang & Ying, 2009), it is important to promote ways that practitioners can support the inferences from these test scores.

Purpose

Many methods for examining person fit in the context of paper-pencil tests exist (Karabatsos, 2003; Ludlow, 1986; Meijer & Sijtsma, 2001; Reise, 2000), but the research examining person fit in CAT is relatively sparse (Meijer & van Krimpen-Stoop, 2010; van Krimpen-Stoop & Meijer, 1999, 2000). The purpose of this study is to explore person fit in a computer adaptive test using a two-stage, statistical and graphical procedure. The research question is as follows: *How can person response functions (PRF) in conjunction with person fit statistics detect and inform practitioners of misfit in CAT?*

CAT Overview

In computer adaptive testing, tests are comprised of items that are customized for each test taker. Because student achievement is assumed to be normally distributed in the population, customized tests are appealing because they hold the promise of measuring achievement equally well for individuals located in the tails and in the middle of the achievement distribution. Other potential advantages of CAT include shorter test lengths compared to traditional paper-pencil tests, immediate scoring and reporting, and facilitating individualized learning (Chang, 2015; Weiss, 1982).

Computer adaptive test administration follows a series of rule-based steps or algorithms that dynamically select and administer test items to the test taker, and then compute an achievement level. Several texts have described the complexity of adaptive test algorithms (e.g., van der Linden & Glas, 2010; Wainer, 2000), so these details will be omitted here. However, a quick summary of the three main algorithms, the *start* step, the *continue* step, and the *stop* step, is included.

To start the adaptive test, a single or a set of test items are chosen from the item database. These items are usually, but not always chosen at random. The purpose of the start step is to obtain an initial estimate of the test taker's achievement level from which the iterative testing process can begin. In the continuing step, items are successively chosen from the remaining items in the bank, and the achievement level is re-estimated after each item is administered, using all of the previously administered items. In this step, the next item is selected based on its parameter values, which maximize the information (or minimize the variance) of the current achievement estimate. In practice, non-psychometric item characteristics, such as the curricular content measured by the item and the exposure rate of the item are also used to select the next item for the test taker. The testing procedure stops when a pre-specified criterion has been satisfied, such as after a certain number of items have been administered to the test taker or after a certain level of precision of the achievement estimate has been met. Additional non-psychometric criteria may also be incorporated into the stopping criterion. After the test has been stopped, the final estimate of achievement is calculated. This final estimate is used as the test taker's final score.

Person Fit

According to IRT models, item and person characteristics determine the responses that a person gives to the items on a test. The psychometric problem at the heart of person fit research is whether or not the final score can predict, with some degree of accuracy, the individual responses to the test items. When good person fit to the model is observed, person responses to the test items are consistent with the IRT model chosen to represent the data. When poor person fit, or person misfit to the model is observed, person responses to the test items are not consistent with the IRT model. Person misfit indicates that the test score assigned to the person may not be a reasonable representation of his or her achievement level. Caution should be practiced when using these scores to make educational decisions.

Approaches that evaluate person fit in paper-pencil testing contexts have been well-documented in the measurement literature. Historically, person fit statistics have been the primary method of evaluating person fit, although graphical methods, which create displays of misfit (e.g., Emons, Sijstma, & Meijer, 2004; Trabin & Weiss, 1979), and explanatory methods, which attempt to use person or item variables to model observed misfit (e.g., Reise, 2000), have also been developed. Person fit statistics provide a numerical description of how expected or likely a person response pattern is compared to a criterion, such as what is expected under a measurement model or what is likely given the observed distribution of responses on the test. Extreme values of the person fit statistic, for example values located in the upper and lower 2.5 percent of the statistic distribution, indicate that a person's response pattern is unexpected or unlikely. Person response functions (PRF) are graphical illustrations of the relationship between the difficulty of the items on a test and the probability of a person giving the correct

answer to these items. PRF are a useful tool in person fit research because they can visually depict the location of misfit (Sijtsma & Meijer, 2001), or in other words, to which items the test taker gave unexpected responses.

Person Fit in CAT

Some studies have explored person fit in CAT. Most of these studies used traditional person fit statistics for detecting person misfit in CAT. Researchers have found that the sampling distribution of some traditional person fit statistics in CAT do not hold to their theoretical distributions, which makes the interpretation about the existence of misfit difficult (Glas, Meijer, & van Krimpen-Stoop, 1998; McLeod & Lewis, 1999; Nering, 1997; van Krimpen-Stoop & Meijer, 1999). One hypothesized reason for this problem is that most person fit statistics require a wider range of item difficulty to detect misfit than what is provided in CAT.

McLeod and Lewis (1999), Meijer (2005), and van Krimpen-Stoop and Meijer (2000), have proposed and evaluated CAT-specific person fit statistics. The results have been mixed. Meijer (2005) reported that the detection power of a CAT-specific person fit statistic was higher than the detection power of other person fit methods in CAT, which included traditional person fit statistics. van Krimpen-Stoop and Meijer (2000) reported similar detection rates for their CAT-specific person fit statistic as was found for traditional person fit statistics. McLeod and Lewis (1999) reported that their CAT-specific person fit statistic was not powerful for detecting misfit in CAT.

Some of these same researchers have promoted using a different statistical framework for conceptualizing person fit in CAT because the items on each adaptive test cover a different range of difficulty. Bradlow, Weiss, and Cho (1998) and van Krimpen-

Stoop and Meijer (Meijer & van Krimpen-Stoop, 2010; van Krimpen-Stoop & Meijer, 2000, 2001) introduced the cumulative sum procedure, CUSUM, for detecting person misfit in CAT. The idea behind using CUSUM requires conceptualizing the response data from a computer adaptive test as an ordered collection of data points. CAT items are targeted where the probability of a test taker giving a correct response is around 0.50, so strings of correct or incorrect answers are not expected. The CUSUM procedure is sensitive to strings of positive or negative residuals, so if the cumulative sum is more extreme than user-defined thresholds (positive and negative), then the response vector can be considered misfitting. These researchers have promoted CUSUM as a viable way to detect misfit in CAT.

Person Fit Approach Used in this Study

The previous research exploring person fit in CAT suggests that person misfit may not be well-understood in these tests. Person fit approaches that provide both statistical and graphical information can provide researchers two pieces of information with which to understand person *misfit* in CAT. The CUSUM method seems to be an attractive approach in this respect. Both Bradlow et al. (1998, p. 918) and Meijer (2002) used plots of the CUSUM procedure to enhance the interpretation of person misfit in CAT. By plotting the CUSUM results, these researchers illustrated where along the latent variable continuum misfit was found and why the response pattern was flagged by the CUSUM procedure (e.g., because the number of correct responses was higher than what was expected by the model).

More research that explores *how* person fit statistics detect misfit in CAT or research that provides supplemental information regarding misfit detection in CAT are

needed. In this study, a different approach that provides statistical and graphical information was used to explore person fit in CAT. Person fit statistics were used to statistically quantify the misfit of person response vectors. Then, person response functions (expected and observed) were used to graphically depict misfitting person response vectors. This general approach to examining fit has been explored using paper-pencil tests (Emons, Sijstma, & Meijer, 2005; Nering & Meijer, 1998; Perkins, Quaynor, & Engelhard, 2011; Ferrando, 2014; Walker & Engelhard, in press). It seems extendable and useful for exploring person fit in CAT.

Conceptual Framework

Responses on computer adaptive tests are stochastic, but from a model-data fit perspective, it is reasonable to expect that the person response vector should accord with the IRT model chosen to calibrate the items and estimate the final score. This idea is shown in Figure 10. Person fit statistics that are designed to gauge the fit of a response vector without reference to the response vectors given by other test takers are theoretically compatible for use in CAT because in CAT there is a limited number of persons who take similar test items or the same collection of test items. Person response functions provide a way to visually inspect the fit of a response vector to an IRT model and can provide supplementary evidence that misfit exists. The expected shape of the response functions (based on what the model dictates) is compared to the shape of the response function that underlies the person's actual responses (i.e., observed in the response data). The amount of discrepancy between the expected and observed functions is an indication of person *misfit*.

There are several fit person fit statistics that could have been chosen to detect person misfit for this application (Karabatsos, 2003; Meijer & Sijtsma, 2001). It seemed reasonable to allow the choice of IRT model used for the creation of the item characteristic curves and the possible types of misfit that may occur in CAT to inform the choice of the person fit statistics. In computer adaptive tests, each person can potentially receive a different set of items with which his or her achievement level is to be estimated. The Rasch model (1960/1980) is theoretically compatible with this test procedure because of its property of invariance, where the particular subset of items that are used to estimate the person's achievement level does not influence the outcome measure (Wright & Stone, 1979). This definition of invariance is unique to the Rasch model because under the two- and three-parameter models, the particular items that are used to estimate the person's achievement level influence the outcome measures.

According to the Rasch model, the probability of a person correctly answering a dichotomously scored item (where 1 denotes a correct response and 0 denotes an incorrect response) is

$$\Pr(X_i = 1 | \theta_n, \delta_i) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}, \quad [13]$$

where θ_n represents the achievement level of person n on the latent variable and δ_i represents the difficulty level of item i on the latent variable.

Three person fit statistics designed to detect misfit to the Rasch model were chosen for this application: Outfit MSE (Wright & Stone, 1979), Infit MSE (Wright & Masters, 1982), and Between fit MSE, Bfit (Smith, 1985). The Outfit and Infit person fit statistics are good for detecting random disturbances in measurement, such as what may be produced by random guessing behavior or careless responding (Smith & Plackner,

2010). The Bfit person fit statistic is good for detecting more systematic disturbances in measurement, such as what may be produced by persons who run out of time, are sub-experts or have a deficiency in a content domain, or are slow to warm-up to the test (Smith & Plackner, 2010). Because these types of responding can potentially occur in adaptive tests, these fit statistics seemed a good match for this analysis.

With respect to person response functions, the shape of Rasch PRF in CAT is not likely to follow the ogive-shape that is customarily seen in paper-pencil tests. This is primarily due to the fact that the range of item difficulty of CAT is narrower than the range of item difficulty of paper-pencil tests. As a result, the shape of a CAT PRF may be more linear than a paper-pencil PRF. This linear shape, however, can prove useful for contrasting fitting and misfitting response patterns. For instance, when an expected response function is superimposed on an observed function (i.e., one that represents the underlying pattern in a test taker's responses), the discrepancies between what is observed and what is expected can be clearly seen.

Method

Study Design

In this study, I used simulated data and visually examined person response functions of two groups of CAT examinees: those examinees whose responses fit the model and those whose responses did not fit the model. First, a separate simulation study was conducted to establish threshold values for categorizing a response vector as fitting or misfitting the model for the three person fit statistics. Next, these threshold values were applied to the 5000 CAT examinees to classify their response vectors as fitting or misfitting the Rasch model. Lastly, person response functions from examinees belonging

to the misfitting (and fitting) categories were plotted and visually examined for discrepancies between what was expected by the model and what was observed in the response vector. It was hypothesized that the person response functions of persons who fit the model would look characteristically different from the person response functions of persons who did not fit the model. It was also expected that the different person fit statistics used in this study would capitalize on different unexpected responses and that these differences would be visually observed in the person response functions.

Data Generation

In operational testing situations, some person misfit is expected to exist, but the amount and type of person misfit is unknown. For the exploratory analysis planned in this study, the amount of misfit needed to be controlled, but the type of misfit did not. To produce this scenario, adaptive test data simulated to fit the model were generated.

Five hundred items were generated to represent a unidimensional item bank calibrated with the Rasch model using the catR package for the R platform (Magis & Raiche, 2012). These items were uniformly distributed over the logit range of -5 to 5. Using the same catR package (and the 500 items), an adaptive test administration was simulated. Specifically, dichotomous item responses for 5000 examinees drawn from a standard normal distribution, $\theta \sim N(0,1)$, were generated. To simulate a dichotomous item response, a random number from a uniform distribution, $U(0,1)$ was compared to the probability of giving the correct response computed from the Rasch model (Equation 13). When the random number exceeded the probability of giving the correct response, the response was set to 1; otherwise, the response was set to 0.

Achievement level estimation was calculated with maximum likelihood procedures. A new provisional achievement estimate was computed after every response starting after three randomly selected items with approximate difficulties of -2, 0, and 2, were administered. The next item (i.e., item 4 and beyond) was selected for the test taker based on the item's proximity to her current provisional estimate of achievement. This item selection process is the same as maximum information selection (Thissen & Mislevy, 2000) when the Rasch model is used (Magis & Raiche, 2012). No content coverage or item exposure constraints were placed on the item selection.

The test was stopped after 40 items were administered. Although forty items may be considered to be fairly long for CAT, forty items represent a typical length of large scale educational achievement tests in practice. Moreover, forty items provides the achievement estimator time to make-up for a poor start (van der Linden & Pashley, 2010). The final achievement level estimation was also calculated using maximum likelihood estimation.

As for the choice of achievement estimator, researchers have suggested that maximum likelihood (ML) estimators are accurate and efficient for moderate to long CAT tests (i.e., those longer than 30 items) that are designed using the Rasch model (Chang & Ying, 2009). Other researchers, however, have criticized maximum likelihood estimation for providing biased achievement estimates when the number of items is small. In this study, the first several provisional achievement estimates were calculated after a small number of items were answered. To check that the ML achievement estimator did not severely bias the CAT results, I also performed the CAT analyses using

the weighted maximum likelihood estimator (Warm, 1989), which has been shown to correct for estimation bias when the number of items is small.

Data Analysis

The primary focus of this study was to examine person fit in the context of a computer adaptive test. All analyses for examining person fit were conducted using the final achievement estimates yielded from the CAT procedure. The three person fit statistics, Outfit, Infit, and Bfit were computed for each simulated examinee using the final achievement estimate, the item difficulty values, and the dichotomously scored item responses. The expected probabilities for the 40 items used to plot the expected person response functions were calculated with the final achievement estimates and item difficulty values. The probabilities used to plot the observed person response functions were calculated using the simulated dichotomous responses, 0 and 1.

Accurate person fit detection relies on acceptable estimation of achievement levels. To gauge the overall adequacy of the CAT simulation process for producing acceptable person estimates, the mean bias, mean absolute bias (MAB), and the relationship between the true and estimated achievement levels were evaluated before the person fit analyses were conducted.

Mean bias was defined as

$$\frac{\sum_{i=1}^N (\hat{\theta}_i - \theta_i)}{N} . \quad [14]$$

It provides an indication of the direction of the estimation error. The closer this value is to 0, the better the estimates are assumed to be because the observed errors are centered and symmetrical around 0.

Mean Absolute Bias was defined as

$$\frac{\sum_{i=1}^N |\hat{\theta}_i - \theta_i|}{N} . \quad [15]$$

In calculating the MAB, the absolute values of the residuals are averaged. This provides an indication of the average magnitude of error in the estimates. The closer the MAB value is to 0, the better the estimates are assumed to be because less error is introduced.

The relationship between the true and estimated achievement levels was described with the Pearson product moment correlation coefficient, r . The stronger the correlation, the better the estimates are assumed to be.

Outfit Person Fit Statistic

Outfit MSE looks for general misfit over the entire response pattern. It takes the residuals (the observed responses minus the expected probabilities of giving correct responses), squares and standardizes them, then sums them over all the items:

$$\text{Outfit MSE}_n = \frac{1}{L} \sum_{i=1}^L \frac{(X_{ni} - E_{ni})^2}{V_{ni}} . \quad [16]$$

In this formulation, X_{ni} is the observed response for person n on item i (either 0 or 1), E_{ni} is the expected response for person n on item i based on the estimated achievement level for the test (a probability calculated from the model), V_{ni} is the variance, i.e., $E_{ni}(1-E_{ni})$, and L is the number of items on the test (Smith, 1985).

Outfit has been used extensively in Rasch analyses for paper-pencil test contexts (Magis, Beland, Raiche, 2014; Petridou & Williams, 2007). Research has shown that when Outfit is applied to items (i.e., item fit analysis), it is sometimes influenced severely by small numbers of unexpected responses at the tails of the achievement distribution, or

in other words, when persons with high estimated achievement earn incorrect answers to easy items or persons with low estimated achievement earn correct answers to difficult items (de Ayala, 2009). Less research has been conducted using Outfit applied to persons (i.e., person fit analysis), but in principle the same problem could exist. Generally speaking, there are fewer items used to measure persons than persons to calibrate items, so unexpected performance at the tails of the item difficulty distribution may influence the Outfit person fit statistic severely. The Infit statistic was developed to reduce the effect of extreme outliers on the Outfit fit statistic (Wright & Masters, 1982).

Infit Person Fit Statistic

Infit MSE statistic also looks for general misfit over the entire response pattern, but in the calculation of Outfit the squared standardized residuals are weighted by the variance (Smith, 1991):

$$\text{Infit MSE}_n = \frac{\sum_{i=1}^L (X_{ni} - E_{ni})^2}{\sum_{i=1}^L V_{ni}} . \quad [17]$$

In this formulation, the terms are the same as in the formulation of the Outfit statistic.

Specifically, X_{ni} is the observed response for person n on item i (either 0 or 1), E_{ni} is the expected response, V_{ni} is the variance, and L is the number of items on the test (Smith, 1991).

The effect of the weighting is that the residuals closest to the estimated achievement level are given more influence in the computation of the fit statistic than the residuals farther from the estimated achievement level. In CAT, the goal is for most of

the items to target the person's achievement level, so the idea of using a fit statistic that weights the residuals is conceptually fitting.

Between Fit (Bfit) Person Fit Statistic

The last person fit statistic used in this study was the Bfit statistic. In the Rasch model, it is assumed that persons' achievement estimates for the total test should predict their observed scores on different subsets of test items (Wright & Stone, 1979). The Bfit statistic tests the tenability of this assumption by comparing expected scores on different subsets of test items (Smith, 1986). A large Bfit value occurs if a person's achievement estimate from her performance on the total test cannot account for her performance on one or more of the item subsets. The item subsets for the Bfit statistic are established a priori, and they can be based on any grouping, such as the order of item presentation, item difficulty, or item content clusters.

The person formulation of the Bfit statistic is

$$\text{Bfit MSE}_n = \frac{1}{(J-1)} \sum_{j=1}^J \frac{\left(\sum_{i \in j}^{L_j} X_{ni} - \sum_{i \in j}^{L_j} E_{ni} \right)^2}{\sum_{i \in j}^{L_j} V_{ni}}. \quad [18]$$

In this formulation, J is the number of item subsets and L_j is the number of items in each subset (Smith, 1985). All other terms are the same as was defined for the Outfit and Infit statistics. The residuals of different item subsets are each summed, squared, and then standardized. Finally, these item subset values are combined to obtain one statistic per person.

In this study, Bfit was computed based on item administration order. Person responses to three groups of items were compared: items 4-15 ($n=12$), items 16-27

($n=12$), and items 28-40 ($n=13$). The first 3 items were omitted from the Bfit analysis since they were used to start the computer adaptive test algorithm. In the context of this study, a large Bfit value would suggest that performance on the first, middle, and last groups of items is the cause of person misfit, rather than general misfitting responses over the entire response pattern, such as what may be detected by the Outfit or Infit person fit statistics. The values for the three fit statistics can range from 0 to ∞ and are assumed to approximate a chi-square distribution (Wright & Stone, 1979; Smith, 1991; Smith & Hedges, 1992). The expected values of Outfit, Infit, and Bfit are 1 when the data fit the Rasch model.

Two general types of response patterns are discordant with the Rasch model: muted response patterns and noisy response patterns. Noisy response patterns indicate that the response data are too unruly to be governed by the stochastic model, or in other words, that the model alone cannot adequately account for the responses (Engelhard, 2013a). Values greater than 1 signify noisy response patterns. Substantively, noisy response patterns may indicate random responding or person dimensionality rather than the unidimensionality that the model assumes.

Values less than 1 signify less variation in the response pattern than what is expected by the model. Muted response patterns indicate that the fit of the responses to the model is too good to be true, and this suggests some dependency in the responses. Substantively, muted responses may indicate item exposure (cheating) or very slow, methodical responding.

In most person fit research, the concern is with identifying noisy response patterns (Reise & Due, 1991), or those patterns with values substantially higher than 1. This

concern was the focus of this study as well. The term misfitting referred to extreme person fit values located in the upper tail of the person fit statistic distribution.

Categorizing Person Misfit

With the simulated design that was used, widespread person misfit was not expected, but the number of response vectors that truly misfit the Rasch model was unknown. A method was needed to classify the person response vectors as fitting or not fitting the model. This is usually done by establishing a person fit statistic threshold value and using it to categorize the individual person fit statistic values as fitting or misfitting.

Methods for establishing a threshold value requires an element of subjective judgment (Drasgow, Levine, & Zickar, 1996). One procedure that has been used and recommended in the recent literature uses a fixed α (i.e., Type I error rate) and derives the critical threshold by simulation (van Krimpen-Stoop & Meijer, 2000; Lamprianou, 2013; Petridou & Williams, 2007; Seo & Weiss, 2013). The rationale for using this procedure is that even though response data are generated to fit a model, there will still be some randomness in the data. A Type I error rate (i.e., rate of misclassifying a fitting response vector as misfitting) is set, and the value of the fit statistic that is located at this point along the distribution of observed fit statistics is used as the threshold. Individual person fit statistics that are more extreme than the threshold are defined as misfitting and values that are less extreme are defined as fitting. A commonly used Type I error rate for establishing person fit statistic threshold values is the conventional $\alpha=0.05$ level (van Krimpen-Stoop & Meijer, 2000; Lamprianou, 2013; Petridou & Williams, 2007), although other α may be used (e.g., Cui & Leighton, 2009).

In this study, the process for establishing the threshold values for the person fit statistics followed a similar process. Five computer adaptive data sets with 10,000 fitting item response vectors were constructed for persons with true achievement levels of -2, -1, 0, 1, and 2 (i.e., 2000 response vectors per achievement level) by using five CAT item banks that were constructed using the same specifications as for the first item bank. This replication of CAT data allowed for an examination of the stability of the threshold values across achievement levels.

Person fit statistics were computed for all simulated test takers. The value of the person fit statistic at the 95th percentile (i.e., Type I error rate of $\alpha=0.05$, one-tailed) was identified as the threshold score for each achievement level cluster. Next, the threshold values for Outfit, Infit, and Bfit were calculated as a weighted average of the five achievement clusters with weights that represented a normal distribution of achievement in the population (van Krimpen-Stoop & Meijer, 2000). Specifically, the fit statistic value at the 95th percentile for achievement level 0 was weighted with 0.50, the value at the 95th percentile for achievement levels -1 and 1 were weighted with 0.20, and value at the 95th percentile for achievement levels -2 and 2 were weighted with 0.05.

Person Response Functions

In this study, the focus was on exploring a two-step procedure using person fit statistics and person response functions for detecting person *misfit* in CAT. Fourteen person response functions were created for persons whose response patterns did not fit the Rasch model according to their fit statistics. These 14 were chosen out of the total number misfitting persons (N=582) and were chosen to represent a range of person achievement levels (approximately -2 through +2 logits) and person statistics flagged

(Outfit, Infit, and Bfit). They were also chosen to represent the more extreme values of Outfit, Infit, and Bfit that were observed. The rationale for this decision was that all of the response vectors in this study were simulated to fit the Rasch model, so by choosing the vectors that produced the more extreme fit values I was choosing those that were likely to truly misfit the model.

For purposes of visual comparison, person response functions were also created for 11 selected persons whose response patterns fit the Rasch model adequately (i.e., they were not flagged by Outfit, Infit, or Bfit). Thirty persons were selected at random from the 4418 persons that had fitting response vectors. Out of these 30, I selected 11 that had different person achievement estimates.

Expected and observed person response functions were created for these 25 examinees. The expected PRF were created by plotting the expected Rasch probabilities. These were computed by inserting the final achievement estimate into Equation 13 as θ_n and the item difficulty parameters as δ_i . The observed PRF were created by smoothing the test taker's original dichotomous response vector. Using a Hanning sequence (Velleman & Hoaglin, 1981), the dichotomous responses to the items (which were first ordered by item difficulty) were transformed to continuous values between 0 and 1.

Hanning can be thought of as a uniform kernel smooth:

$$s_i = (y_{i-1} + 2y_i + y_{i+1}) / 4. \quad [19]$$

The first and last responses (i.e., y_1 and y_n) are left as-is. For the responses to items y_2 through y_{n-1} (items 2 through 39 in this study), s_i replaces the observed responses, y_i . The response y_i receives a weight of two and the two responses adjacent to y_i on each end receive a weight of 1.

In this study, the goal of the smoothing procedure was to obtain an adequate smooth, so that the final PRF was useful, while preserving enough of the original response pattern, so that the final PRF remained truthful. To achieve this goal with dichotomous data, the Hanning sequence was repeated a number of times. Following Engelhard (2013b), I repeated the Hanning sequence one time for each raw score point. That is, a person who obtained a raw score of 20 out of 40 had her dichotomous response vector smoothed 20 times.

The final smoothed values were plotted on the same coordinate space as the expected probabilities. Thus, for each person ($N=25$), two PRF were constructed. The visual inspection of the PRF focused on the discrepancy between the expected function and the observed function. Because the PRF were not dependent on the fit statistics chosen to detect misfit, this two-step procedure was equivalent to providing two independent ways of examining person fit.

Results

Evaluation of the Computer Adaptive Test Procedure

Before the person fit analyses commenced, the estimated achievement levels from the adaptive test procedure were evaluated for accuracy. Specifically, two sets of simulated results were evaluated: one that used maximum likelihood estimation for estimating achievement and one that used weighted maximum likelihood estimation. The difference between the estimated and true achievement levels using these two methods were negligible, with the mean bias, mean absolute bias, and Pearson product moment correlation being discrepant at the thousandth decimal place or smaller. To avoid

redundancy, only the results obtained from the maximum likelihood procedure are presented.

The mean bias was 0.001 and the mean absolute bias was 0.268, which was considerably smaller than the average standard error of the estimate for the sample (mean SEM = 0.331). The Pearson product moment correlation coefficient for the estimated achievement levels and the true theta levels was 0.95, $p < 0.001$. In terms of precision, 67% of the true achievement levels fell between one standard error of the achievement level estimates (i.e., ± 1 SEM) and 95% of the true achievement levels fell between two standard errors of the achievement estimates (i.e., ± 2 SEM). Taken together these results suggest that the CAT procedure produced adequate estimates of achievement.

Categorizing Person Misfit

The first step of the person fit analysis required establishing threshold statistics for Outfit, Infit, and Bfit to be used for categorizing the response vectors as fitting or misfitting. Table 4 shows the results of the simulated replications of CAT data used for this purpose. In Table 4, the mean value at the 95th percentile by achievement level cluster across the five simulated datasets for the Outfit, Infit, and Bfit statistics are presented. The values reported in Table 4 indicated that the person fit statistic values at the 95th percentile were independent of achievement level, and suggested that using a single threshold value (one for each fit statistic) to categorize response vectors as fitting or misfitting was justified.

The values included in Table 4 were used to create a weighted average. The resulting threshold values were 1.166, 1.050, and 2.997 for Outfit, Infit, and Bfit, respectively. These values are included at the bottom of Table 4.

Person Fit

For the person fit analysis, the three person fit statistics were computed for every simulated test taker (N=5000). Then, using the misfit threshold values at the bottom of Table 4, each test taker was categorized as either fitting or misfitting the model three times, once for each person fit statistic. Person fit values greater than the threshold were defined as misfitting and person fit values less than the threshold were defined as fitting.

The percentages of test takers flagged for misfit for each fit statistic was 4.6% for Outfit, 4.3% for Infit, and 5.4% for Bfit. Out of 5000 test takers, 582 (11.6%) were flagged by at least one of the three person fit statistics. Out of this, 123 (21.1%) were flagged by both Outfit and Infit. Less than 2% of this 582 were flagged by both Outfit and Bfit (1.0%) and both Infit and Bfit (1.5%). Only 3 response vectors out of the 582 were flagged by all three fit statistics (0.0%). The absence of overlap in test takers flagged by Bfit and Outfit (and Bfit and Infit) suggested that the Bfit fit statistic captured a different type of misfit than Outfit and Infit. The moderate overlap between Outfit and Infit suggested that Outfit and Infit capture similar, but not exactly the same type of misfit.

Person Response Functions

Person response functions provided a way to visually inspect a person's response vector and in this study were used to corroborate the statistical judgment regarding misfit. The person response functions of 14 misfitting test takers were created, organized by type of misfit (Outfit, Infit, and Bfit), and evaluated visually. All 14 PRF were distinctive, however, similar characteristics were noticed. These similarities were explored further by contrasting the 14 *misfitting* PRF with PRF from 11 fitting response vectors. Table 5

includes information about the 25 test takers who were chosen for the graphical step of the analysis. Figures 11-13 show 20 PRF that were selected to illustrate the characteristics that were observed in the 25 PRF.

For the PRF plots in Figures 11-13, the items are ordered by difficulty and are located on the x-axis. The probability of giving the correct response is located on the y-axis. A reference line is included at $\Pr(x=1)=0.50$, the location of the achievement estimate in the Rasch model. The dichotomous responses are represented by asterisks, and a square represents the final achievement estimate from the CAT for the person.

The solid circle function represents the Rasch expected person response function. The SEM bands are represented by the dotted lines. These are calculated by plotting the Rasch probabilities for the estimate achievement level (θ) plus 2 standard errors and minus 2 standard errors, i.e., plotting $\Pr(x=1|\hat{\theta} \pm 2*SEM)$. Lastly, the hashed diamond function represents the observed (smoothed dichotomous) person response function.

Figure 11 shows misfit in the computer adaptive test as detected by Outfit. In Figure 11, the first column of PRF (PRF 938 and 1254) shows persons whose response vectors fit the model. The last two columns of PRF (PRF 4417, 3650, and 750) show persons whose response vectors misfit the model. One common observation for the persons flagged as misfitting that is illustrated in Figure 11 is the large unexpected correct (or incorrect) response at the end (or beginning) of the response vector. For instance, in PRF 4417, the hashed diamond located at item difficulty -2 illustrates that this person gave an incorrect answer to this item. The solid circle located at item difficulty -2 illustrates that the model expected the person to give a correct answer to this item. The other two misfitting PRF included in Figure 11 (PRF 3650 and 750) show a similar

discrepancy. This commonality suggests that the Outfit statistic in CAT is sensitive to unexpected responses to items located at the ends of the item difficulty continuum.

However, a single unexpected response at the ends of the item difficulty continuum did not appear to be enough to trigger Outfit misfit. PRF 1254, which was categorized as fitting the model, also exhibits an extreme unexpected response at 2 logits. But, this PRF illustrates the second common observation for the persons flagged as misfitting by Outfit: the moderate unexpected responses (shown as jitter) in the middle of the response vector. For the misfitting PRF, there are some peaks and dips in the observed response function located in the middle of the item difficulty continuum or near the estimated achievement level in addition to the unexpected response at the ends (i.e., an extreme unexpected response). For the fitting PRF (PRF 938 and 1254), either peaks and dips in the middle, *or* an extreme unexpected response are present, but not both. Thus, it appears that when a single unexpected and extreme response is combined with moderate misfit throughout the response pattern in CAT, Outfit detects it.

Figure 12 shows misfit in the computer adaptive test as detected by Infit. In Figure 12, the first column of PRF (PRF 4403 and 2689) show persons whose response vectors fit the model. The last two columns of PRF (PRF 4327, 3638, 4494, and 832) show persons whose response vectors misfit the model. The common observation for the persons flagged as misfitting that is illustrated in Figure 12 is the wave-like jitter around the probability of 0.50 or the estimated achievement level. For instance, in PRF 3638, the hashed diamonds show that between item difficulty -2 and -0.25, the observed probabilities of giving a correct answer do not accord with the expected probabilities of giving the correct answer. Instead of a steady decline in the probabilities (shown by the

solid circles), dips and peaks between the probabilities of .2 and .6 are shown. The other three misfitting PRF included in Figure 12 (PRF 4327, 4494, and 832) show a similar pattern. This commonality indicates that the Infit statistic in CAT is sensitive to unexpected responses to items located near the estimated achievement level (or where the probability of giving a correct response is near 0.50).

It appeared that the amount of unexpected responses that were necessary to trigger misfit depended on the range of item difficulties that the unexpected responses spanned. PRF 2689, which was categorized as fitting the model, also exhibits some unexpected responses near the probability of 0.50. The main difference that can be observed between this PRF and PRF 4327, which was categorized as misfitting the model, is the range of item difficulties covered by the misfit. For PRF 4327, the wave-like jitter covers logits -1 to 1.5, a range of 2.5 logits. For PRF 2689, the wave-like jitter covers logits 2 to 3, a range of 1 logit. If the peaks and dips are moderate and cover a wide range of item difficulties, the Infit statistic becomes large.

Figure 13 shows misfit in the computer adaptive test as detected by Bfit. All three PRF included in the top row of Figure 13 misfit the model. However from the inspection of these PRF, no obvious or consistent nuance of the PRF emerged that could illuminate how or why these response vectors misfit the model. For PRF 2227, the observed function (i.e., the hashed diamonds) looked like they matched the expected function (i.e., the solid circles) well. The observed PRF for 1431 and 4556 looked vertical (i.e., Guttman-like) compared to the monotonic expected PRF, but overall no unexpected responses are seen from the plotted function. Moreover, the Infit and Outfit values for

these persons (i.e., 1431 and 4556 from Table 5) are less than 1, which would describe response patterns that are *muted*, not noisy.

The person response function analysis for the Bfit statistic was not completely satisfying because I was unable to ascertain why the Bfit statistic was so large. In an effort to examine these response vectors in a way that may help illustrate the misfit detected by Bfit, residual plots were created for these persons. These plots are included on the bottom row of Figure 13. In the residual plots, the items were placed *in the order in which they were administered* on the x-axis, and the standardized residuals were placed on the y-axis. This new plot configuration was chosen because the Bfit statistic looks for differential person fit across item subsets. In this study, the three item subsets used for calculating the Bfit statistic were defined by their location in the test (i.e., first subset, middle subset, and last subset). In the residual plots at the bottom of Figure 13, three rectangular boxes are drawn to illustrate the three item subsets included in the Bfit analysis. It was expected that within each PRF, different patterns of residuals across the three item sets would be observed, and that this could help illustrate how Bfit detects misfit in CAT.

For residual plot 2227, this explanation appeared to be accurate. The standardized residuals in the first subset of items exhibited a diagonal (sloped) pattern instead of the horizontal pattern that was exhibited in the second and third subsets of items. For residual plots 1431 and 4456, however, this explanation did not hold. For these persons, no clear difference between the residuals across the three subsets of items could be determined from the plots. As a result, the implications for how the Bfit statistic detects misfit in CAT were inconclusive.

Discussion

Not all test scores are equally trustworthy for representing what students know and can do (and what they should learn next). Scores of persons who provide response vectors that do not fit the model are not justifiable or trustworthy. In practice, it is hoped that after proper attention has been paid to item development and quality checks have taken place, most student responses will fit the model, while only a few will not. This ideal scenario was simulated in this study with the aim of exploring person fit in a computer adaptive test. First, computer adaptive test data were simulated to fit the Rasch model. Then, three person fit statistics were used to categorize person response vectors as fitting or misfitting the model. Lastly, person response vectors of misfitting persons were visually inspected using person response functions with a focus on identifying discrepancies between the patterns of the expected and observed response functions.

The computer adaptive test yielded achievement level estimates that were satisfactory overall. However, being flagged for individual misfit meant that the *individual* achievement estimate and the difficulty level of the items alone could not predict the person's responses to the test items (noisy misfit). From this exploration, several findings emerged that can help to understand person *misfit* in CAT.

The research question asked how person fit statistics and person response functions together could aid practitioners in detecting model-data misfit in the context of CAT. The results illustrated that the Outfit statistic detected misfit at the ends of the item difficulty continuum in CAT. This is similar to the way that Outfit is known to detect misfit in paper-pencil tests (Wright & Masters, 1982), although in a well-implemented

CAT, the ends of the item difficulty continuum will cover a smaller range than what is covered in paper-pencil tests.

Yet, more than a single unexpected and extreme response (e.g., a person with a low achievement level gives the correct response to a difficult item) was needed to flag Outfit misfit in the adaptive test. It appears that moderately noisy (underfitting) response patterns with a single extreme unexpected response earns a large Outfit value indicative of person misfit. But neither mildly noisy response patterns with an extreme unexpected response, nor moderately noisy response patterns with no extreme unexpected response, earns an Outfit value indicative person misfit. The implication here is that Outfit may detect misfit a little differently in CAT than in paper-pencil tests, perhaps because generally CAT tests are better targeted than paper-pencil tests. More research, for example research that manipulates item selection procedures such as the wider/narrower targeting of test items in CAT, is needed to support the generalizability of this finding.

As for Infit, the results illustrated that Infit detected misfit where the probability of giving the correct response was near 0.50 in CAT. Detecting measurement disturbances around this location is what Infit is designed to do (Wright & Masters, 1982). However, it appeared that the severity of the misfit needed to be of a certain magnitude before the Infit statistic was large enough to flag the response vector as misfitting. This finding suggests that the sensitivity of the Infit statistic may be suitable for use in computer adaptive tests because most of the items on a computer adaptive test will hover around a probability of 0.50. An easily triggered fit statistic (i.e., one that is too sensitive to stochastic jitter) or a blunt fit statistic (i.e., one that is not sensitive enough to stochastic jitter) would be useless or ineffective.

For Bfit, the results from the two-step procedure did not provide clear evidence for how the Bfit statistic detected misfit in CAT. Although large Bfit values were calculated from some of the simulated response vectors, the PRF and the additional residual plots of the response vectors did not reveal clear and consistent patterns of misfit. It is likely that the design of the study was not conducive to creating the type of person misfit that would be best detected by Bfit. Under the design of this study, random disturbances to the Rasch model were expected. The Bfit statistic looks for systematic disturbances such as what may be exhibited if a person has sub-expert knowledge on a set of items within the whole (like on the geometry items within a 5th grade Mathematics test), or if a person becomes fatigued at the end of a test (Smith, 1985; Smith & Plackner, 2010). It is likely that many of persons who exhibited Bfit misfit in this study are Type 1 errors. More research is needed to examine how the Bfit statistic detects misfit in CAT.

From this exploratory analysis, the Outfit and Infit person fit statistics appear to be promising indices for detecting person misfit in CAT. These person fit statistics when used together with person response functions can assist practitioners and other educational stakeholders in the interpretation of person misfit. The findings reported here are tentative because misfit was not generated in this study, and the accuracy of person fit detection could not be assessed. Future research should manufacture misfit in a simulated CAT setting and evaluate how accurately Outfit, Infit, and Bfit detects it (e.g., McLeod & Lewis, 1998). These findings do, however, highlight the advantage of the two-step, statistical and graphical, procedure for examining person fit. Namely, that this procedure allows for a statistical judgment about person fit of a response vector to be further informed by a visual inspection of that response vector. Given the skepticism

regarding the use of existing person fit statistics in an adaptive test context (Glas et al., 1998; McLeod & Lewis, 1999; Nering, 1997; van Krimpen-Stoop & Meijer, 1999), this additional information is warranted not only for corroborating judgments regarding the validity of a score, but also for deepening our current understanding of how fit statistics detect misfit in these tests.

A last observation that was noted in the study pertained to the establishment of thresholds for categorizing person misfit. The thresholds for Outfit, Infit, and Bfit established for this study were empirically-derived using repeated simulations of adaptive test data and a Type I error rate of 0.05, one-tailed. The threshold value used for Outfit was 1.17, the threshold value used for Infit was 1.05, and the threshold value used for Bfit was 3.00.

The thresholds for Outfit and Infit are close to 1, which is the expected (mean) value of these statistics when the data fit the model, and suggests that the range of Outfit and Infit values that indicate acceptable fit to the model may be smaller for CAT than for paper-pencil tests. As a comparison, consider that a one-tailed, rule-of-thumb interpretation of Outfit and Infit that is based on experience with paper-pencil tests is that values larger than 1.20 (Rudner & Wright, 1995) or 1.30 (Engelhard, 2009; Linacre, 1997) indicate misfit to the model.

The implication is that if person fit thresholds that are established for paper-pencil tests are used to detect misfit in CAT, practitioners may under-detect misfit in CAT. This may be especially true when Infit is used. Previous model-data fit researchers have recommended computing and using empirically-derived threshold values for each different person fit application because the rule-of-thumb interpretation of fit statistics are

not always accurate (Reise, 1990; Seo & Weiss, 2013; Smith, Schumacker, & Bush, 1998; Wu & Adams, 2013). This recommendation seems especially prudent for CAT.

Conclusion

Although achievement estimates from IRT models have been shown to be fairly robust to model-data misfit in paper-pencil tests (Adams & Wright, 1994; Sinharay & Haberman, 2014) and CAT (Glas et al., 1998), persons may respond to items in unique and unstudied ways that bias their achievement estimates in real testing situations. In educational CAT testing where the item parameters are considered known, and where each student may receive a different test form, evaluating individual person fit can provide information about model-data fit. Person fit statistics and person response functions provided complementary information regarding person fit in this study. Methods such as these have a place in adaptive testing quality checks and score reporting because they enhance validity evidence for adaptive test score interpretation and use (APA/AERA/NCME, 2014). Person response functions supplement statistical person fit information by providing visual representations of misfitting patterns. These visual representations can help practitioners *see* person misfit and help provide a substantive meaning or interpretation of person fit in computer adaptive tests.

Chapter Six: Discussion

This chapter reviews the aims and purpose of this research and includes a general discussion of the research findings.

In the *Standards for Educational and Psychological Testing* (APA/AERA/NCME, 2014), it is written that testing professionals should evaluate the validity of claims derived from an educational or psychological test regarding test score meaning and use. This dissertation focuses on one way to inform the meaning and use of educational test scores. Specifically, it is argued that the validity of claims regarding test score meaning and use can be evaluated by person fit, which is a subtype of model-data fit. The primary method used for exploring person fit in this research is the person response function.

The first research question asked *how person response functions and model-data fit contribute to the validation of inferences regarding the meaning of person scores?* This question was answered by examining the literature. The answer is summarized and discussed in the next paragraphs.

Measurement theory, and in this research, item response theory, requires that certain psychometric properties be observed in test data before the resulting test scores can be considered to be trustworthy measures of a construct (Swaminathan et al., 2007). These properties are set forth by the IRT model chosen to govern the data, and the extent to which the observed data shows good fit to the IRT model is the extent to which the scores are trustworthy measures that yield valid inferences. Person fit is a way to examine model-data fit in a set of test data. Stated succinctly, person fit analyses examine how well person responses to the individual items on the test can be predicted by the IRT model and her or his total score. In this dissertation, person response

functions are considered to be tools for visually examining person fit because they illustrate the probability of a person giving the correct response to the individual items. PRF can be created to show the pattern of responses that are *expected* by the IRT model and they can be created to show the pattern of responses that is *observed* in the data. In the Rasch model, the probability of giving the correct response decreases as the items increase in difficulty. The (Rasch) expected person response function follows a decreasing or negative sloping pattern. The observed person response functions typically have no set slope; thus remain free to follow the underlying pattern in the responses. By visually comparing the observed PRF and the expected PRF, researchers can obtain information regarding how well test data fit the Rasch model.

In previous research, statistical approaches, not graphical approaches like those based on person response functions, are most common. But some researchers have argued for using both statistical and graphical methods for evaluating person fit (e.g., Emons et al., 2005; Engelhard, 2013b). The rationale for using multiple ways to evaluate person fit is because the statistical and graphical methods have different strengths.

In reality, person fit exists as a continuous outcome—person responses can range from fitting perfectly to misfitting perfectly and occur at every intermediate step. Person fit statistics provide a framework for detecting *too much* misfit based on traditional statistical theory. The strength of person fit statistics is that they provide a way to make discrete decisions about misfit in a response pattern, like it fits or does not fit the model.

But person fit statistics cannot provide diagnostic or explanatory information about the person responses they detect as misfitting. The strength of person response functions is that they can provide a visual image of person fit and can potentially provide

information about where and in what way person misfit occurs. However with real data that exhibit continuous levels of *misfit* and that hardly ever fit the model perfectly, PRF alone can be somewhat difficult to interpret, especially in terms of making a discrete decision such as whether or not adequate fit is observed. Using person fit statistics and person response functions together combines the strengths of both approaches and provides the benefit of complementary information (de Ayala, 2009).

Based on the implications from previous research on person fit, it appears to be well-known, at least among educational measurement researchers, that person fit evaluation provides important validity information. Yet, person fit procedures are not widely used in educational testing practice (Cui & Roberts, 2013). Moreover, when person fit analyses are conducted, they are typically done for the purpose of checking the stability of item parameters (Embretson & Reise, 2000; Reise & Flannery, 1996) and are conducted over all test takers or groups of test takers, not individually.

Checking item parameter stability is a critical aspect of model-data fit, and it is necessary for establishing the quality of a testing program. Also, it is true that in well-developed and targeted tests most persons can be expected to provide responses that fit the model adequately well (Rudner et al., 1996). But the inferences made on the basis of test scores have implications for *individual* test takers. And although most person responses will fit the model adequately, a few persons will provide responses that do not fit the model adequately. Using global person fit evaluation is not enough to inform whether the inferences about *individual student performance* are justifiable and trustworthy. Individual person fit evaluation can provide evidence for these inferences.

Individual person fit analysis is not only relevant for paper-pencil testing, but for computer adaptive testing as well. In CAT, test takers receive different sets of test items to measure their achievement, and the test algorithms are highly reliant on IRT item parameters for the selection of *customized* items. But because the test items are given at varying rates and given to different test takers in CAT, traditional item stability checks are more difficult to conduct. Individual person fit analyses provide a way to check the quality of person model-data fit in these tests.

In this program of research, I also argued that communicating individual person fit information and its relevance for test score interpretation is essential for the appropriate use of test results. Educational practitioners, policymakers, and other stakeholders should have access to person fit information to use when making educational decisions. As far as I can tell, not much research has been conducted in this area. This missing piece reflects a gap between research and practice. This gap was the impetus for research question two.

Research question two asked *what existing methods of creating person response functions can be utilized in practice for understanding the patterns of person responses and validating the inferences of scores on educational tests?* Three applications included in Chapters Three, Four, and Five explored the idea that person response functions can communicate individual person fit information to educational stakeholders. Intuitively it makes sense, and past researchers have argued that using two sources of person fit information, such as person response functions and person fit statistics, provides more information about the trustworthiness of a test score than one source of information. An argument that I made in this dissertation is that person fit statistics are not easily

understood by educational stakeholders because they do not have a clear and inherent substantive meaning. I reasoned that using person fit statistics in conjunction with person response functions can make idea of person fit meaningful and relevant. With guidance in interpreting the shapes and patterns illustrated in the PRF, practitioners can *see* the implications of person *misfit* for test score use.

In Chapters Three, Four, and Five, a visual comparison between expected and observed person response functions was conducted. Across these three chapters, two methods for creating observed PRF were explored. (In these chapters the expected PRF were created based on fit to the Rasch model.) In Chapter Three, a parametric approach for creating observed person response functions and estimating a person fit index (based on the slope of the PRF) was used. In Chapters Four and Five, a non-parametric approach for creating observed PRF was used. Discrepancies between observed and expected PRF indicated person *misfit* and provided supplemental person fit information to the person fit statistics. In these three applications, person fit procedures were conducted at the individual person level, not at the group level.

The probabilities for the expected response functions in this research were calculated using values obtained from the Rasch model. The probabilities for the observed response functions were calculated using the actual scored test values and were computed with widely available software packages (i.e., HLM, SAS, and R). The formulations that were used in Chapters Four and Five could be computed by persons with little training in educational measurement, although computer programming skill is needed. The goal was to explore a person fit approach that could be used as a final step

for ensuring that an individual test score was a trustworthy representation of what the test-taker knows and can do, and can help decide what she or he should learn next.

In this study, person response functions were used to complement the interpretation of the person fit statistics. And in this study, like real operational tests settings, the amount of absolute or true misfit was not known. The findings suggest that person response functions can be used as a tool to help understand individual person misfit. Yet no single form of person response function is useful for all intended purposes. In Chapter Three, it was demonstrated that the parametric observed person response function provided smooth and monotonic decreasing functions. These functions may be useful for conveying general and substantial misfit, but they cannot convey the nuances of misfit. In Chapters Four and Five, it was demonstrated that the non-parametric observed person response functions provided jagged response functions that conveyed misfit in great detail, but a frame of reference, such as the model-based expected pattern of responses, is needed to best interpret these PRF.

In analyzing the results from Chapters Three, Four, and Five some interesting observations were noted. First, neither a person response function, nor a person fit statistic, appeared as good on its own for understanding individual person misfit as when the procedures were combined. In this research it was noted that persons may obtain the same person fit statistic, but have vastly different person response functions. The information provided by the fit statistics support the conclusion that the response pattern is misfitting in the statistical inference sense. The information provided by the PRF can *show* how the response pattern misfits the model.

Both pieces of information are useful for understanding the person's particular type of misfit and in developing a plan for handling it. For instance, in Chapter Five where a simulated computer adaptive test application was used and in Chapter Four where guessing responses were created, the person response functions showed that for some persons detected by the fit statistics, the probability of a person giving the correct response equaled 0.50 at more than one location along the item difficulty continuum. The discrepancy between the observed and expected PRF illustrated the idea that a single achievement level is not justified given the unruly response data. The implication is that using a single test score for these persons will either under or overestimate their true achievement level and the inferences about what they know and can do may be erroneous.

On the other hand, in Chapter Five, some persons were detected as misfitting by the UB statistic, but in visually examining the PRF, the reason that the response pattern was flagged as misfitting was not clear for some of the persons. Given the incongruence between the person fit statistic and the PRF procedures, further scrutiny of the response pattern is warranted. In these examples, the two pieces of information (statistical and PRF-based) helped with the substantive interpretation of person misfit.

Furthermore, the results from the three applications also highlighted how person fit statistics and person response functions provide different information about individual person fit. Previous researchers have argued that the interpretation of fit statistics depends on the relative amount of misfit that is observed in a particular data set (Wu & Adams, 2013). The range of person fit statistics in Chapters Three, Four, and Five were different, and in Chapters Four and Five it was clear that the range influenced the

threshold value (i.e., the point along the range that was used to determine statistical misfit). In Chapter Five, a person fit statistic value of 1.18 was considered to be misfitting, whereas in Chapter Four this value was not considered misfitting.

By comparison, the person response functions provided information about absolute person fit to the model. PRF showed the match between what is observed in the data and what is expected by the model without regards to the amount of misfit observed in the particular data set. Again, the two pieces of information from this two-step, statistical and PRF-based approach provided useful and informative details for understanding person misfit.

Limitations

There are some limitations that should be considered when drawing inferences based on the results of this research. First, it focuses on only one measurement model, the Rasch model (Rasch 1960/1980). One ramification of choosing a particular model is that the findings may not generalize to similar settings using other measurement models. In the Rasch model, a requirement is invariant item ordering, which is made apparent by equivalent and non-crossing item and person response functions. In measurement models that add additional parameters to better accommodate the unruliness of real test data, unexpected responses are more difficult to detect because the item and person response functions are allowed to cross. In terms of generalizing the interpretations of the PRF explored here to test settings where other measurement models are used, additional research is needed.

A second limitation is that this study focused only on conveying misfit for dichotomously scored test items. In educational tests being developed now (e.g.,

Common Core and Race to the Top assessments), item types that extend beyond right vs. wrong scoring to give credit for partially correct responses will also be included. Person misfit is still a potential problem in these tests, so future research could explore the use of person response functions to examine and communicate misfit for a mix of polytomous and dichotomous items.

One of the contributions of this program of research is that it illustrates a simple and straightforward method for creating and comparing PRF that can be implemented at a local level. It was reasoned that this method *could* clearly communicate person fit information to educational stakeholders. In this research, however, I did not explore if the person fit information provided by the two-step approach was actually understood by stakeholders. More research is needed to evaluate this approach for stakeholder accessibility and use. Mixed-methods or qualitative methodologies seem most suitable for examining these issues.

Significance

A fundamental role of educational measurement is to improve teaching and learning (Miller, Linn, & Gronlund, 2009). But for these improvements to be realized, measurement experts need to understand and communicate the benefits and shortcomings of educational tests. The broad message of this dissertation highlights the idea of person fit and its implications for the meaning, interpretation, and use of test scores (validity).

In this research, it is argued that person fit information should not be under the sole purview of educational researchers and measurement experts; that instead this information should be transmitted to all test score consumers and stakeholders who use test scores to make important educational decisions for students. With this information,

the level of trust that can be placed on the test scores as good representations of student levels of achievement can be used in the decision-making process.

This general message is supported in *The Standards for Educational and Psychological Testing* where it is written that

Validation is the joint responsibility of the test developer and the test user. The test developer is responsible for furnishing relevant evidence and a rationale in support of any test score interpretations for specified uses intended by the developer. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used.” (AERA/APA/NCME, 2014, p. 13).

Person fit statistics and person response functions inform test score meaning and use for individual students. Communicating this information to educational stakeholders who use test data is important for promoting appropriate test score use, but it is not yet being done in educational testing practice. The findings from this program of research present an initial step in making this important goal a reality.

The idea that graphical illustrations may communicate person fit more clearly and comprehensively than person fit statistics, especially to audiences that have limited statistical or technical backgrounds or interests, is persuasive. There is a reason why the old adage about a picture being worth a thousand words is still around today. In mathematics education, where visual and symbolic representation is considered fundamental to mathematical thinking (Janvier, 1987; Kaput, 1987), many researchers focus on the topic of accessibility and effectiveness of illustrations for promoting understanding of difficult or abstract mathematic concepts (e.g., Dufour-Janvier, Bednarz,

& Belanger, 1987; NCTM, 2001). Even in educational measurement research, the use of visual illustrations has been highlighted as a strategy for effectively communicating psychometric concepts to educational policy leaders (Sireci & Forte, 2012).

Similarly, improvements in test score reporting have been witnessed over the past decade and the study of test score reporting has emerged as an important focus for the measurement community (Zenisky, Hambleton, 2012; Zenisky, Hambleton, & Sireci, 2009). On student score reports today, it is common to see the student's test score and a standard error of measure. These two pieces of information assist test score users in interpreting and using the score for pedagogical decisions. Given the interest in creating informative, accessible, and relevant score reports, now may be a sensible time to consider how to incorporate information pertaining to test score *validity*, such as person fit information and person response functions, into test score reporting practices.

References

- Adams, R. J., & Wright, B. D. (1994). When does misfit make a difference? In M. Wilson (Ed.), *Objective Measurement: Theory Into Practice* Volume 2 (pp. 244-270). Norwood, NJ: Ablex.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, D.C.: AERA.
- American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. New York: American Psychological Association.
- Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement*, 10, 67-78.
- Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1-16.
- Anastasi, A. (1988). Validity: Basic concepts. In A. Anastasi (Ed.), *Psychological testing* (pp. 139-164). New York: MacMillan.
- Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care*, 42(1), 7-16.
- Andrich, D. (1989). *Rasch Models for Measurement*. Newbury Park, CA: Sage Publications.
- Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. *Sociological Methodology*, 15, 22-80.

- Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics*, 37, 417-442.
- Baker, F. B. (2001). *The basics of item response theory*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. L. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley Publishing Company.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*, 2nd Edition. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. New York: Cambridge University Press.
- Brunk, H.D. (1981). Estimation of stimulus-response curves by Bayesian least squares. *Psychometrika*, 46, 115-128.
- Bryk, A.S., Raudenbush, S. W., & Congdon, R. (2009). *HLM: Heirarchical linear and nonlinear modeling for Windows, version 6.08* [Computer software]. Chicago: Scientific Software International.
- Buckingham, B. R. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology*, 12,271-275.
- Carroll, J. B. (1983). The difficulty of a test and its factor composition revisited. In H.

- Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederick M. Lord* (pp. 257-282). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Carroll, J. B. (1985). Defining abilities through the person characteristic function. In E.W. Roskam (Ed.), *Measurement and personality assessment* (pp. 121-131). Amsterdam, the Netherlands: Elsevier (North Holland).
- Carroll, J. B. (1990). Estimating item and ability parameters in homogeneous tests with the person characteristic function. *Applied Psychological Measurement*, 14, 109-125.
- Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. New York: Cambridge University Press.
- Carroll, J. B., & Schohan, B. (1953). *Construction of comprehensive achievement examinations for Navy officer candidate programs*. (Research Report NR 154-138). Pennsylvania: American Institute for Research.
- Carroll, J. B., Mead, A., & Johnson, E. S. (1991). Test analysis with the person characteristic function: Implications for defining abilities. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach*. (pp. 109-143). Hillsdale NJ: Erlbaum.
- Chan, E. K. H. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 9-24). Switzerland: Springer.
- Chang, H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*,

80, 1-20.

College Board. (2011a). *About the Test*. Retrieved from

<http://professionals.collegeboard.com/testing/sat>

College Board. (2011b). *Writing assessment*. Retrieved from

<http://professionals.collegeboard.com/testing/sat-reasoning/about/sections/writing>

Conijn, J. M., Emons, W. H. M., van Assen, M. A., & Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis.

Multivariate Behavioral Research, 46, 365-388.

Cronbach, L. J. (1946). Response Sets and Test Validity. *Educational and Psychological Measurement*, 6, 475-494.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, 10, 3-31.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer and H. I. Braun (Eds.), *Test validity*, (pp. 3-17). Hillsdale, NJ: Erlbaum Associates.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, 46(4), 429-449.

Cui, Y., & Roberts, M.R. (2013). Validating student score inferences with person-fit statistic and verbal reports: A person-fit study for cognitive diagnostic assessment. *Educational Measurement: Issues and Practice*, 32(1), 34-42.

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

Guilford Press.

- De Boeck, P., & Wilson, M. (2004). A framework for item response models. In P. De Boeck, and M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (pp. 3-41). New York: Springer-Verlag.
- Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, 9, 47-64.
- Dufour-Janvier, B., Bednarz, N., & Belanger, M. (1987). Pedagogical considerations concerning the problem of representation. In C. Janvier (Ed.), *Problems of Representation in the Teaching and Learning of Mathematics* (pp. 109-122). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Ebel, R. L. (1968). Blind guessing on objective achievement tests. *Journal of Educational Measurement*, 5(4), 321-325.
- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, 36(8), 449-455.
- Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about person-response functions in person-fit analysis. *Multivariate Behavioral Research*, 39, 1-35.
- Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods*, 10, 101-119.
- Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman,

- Rasch, and Mokken. *Measurement: Interdisciplinary Research & Perspective*, 6(3), 155-189.
- Engelhard, G., Jr. (2009). Using item response theory and model-data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement*, 69(4), 585-602.
- Engelhard, G., Jr. (2013a). Hanning (Smoothing) of person response functions. *Rasch Measurement Transactions*, 26(4), 1392-1393.
- Engelhard, G., Jr. (2013b). *Invariant measurement: Using Rasch models in the social behavioral, and health sciences*. New York, NY: Routledge.
- Engelhard, G., Kobrin, J., Wind, S.A., & Chajewski, M. (2014). *Differential item and personfunctioning in large-scale writing assessments within the context of the SAT Reasoning Test*. Research Report. New York, NY: College Board.
- Ferrando, P. J. (2007). Factor-analytic procedures for assessing response pattern scalability. *Multivariate Behavioral Research*, 42, 481-507, doi:10.1080/00273170701382583
- Ferrando, P. J. (2014). A general approach for assessing person fit and person reliability in typical-response measurement. *Applied Psychological Measurement*, 38, 166-183.
- Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, and J.A. Clausen (Eds.), *Measurement and Prediction* (Volume IV, pp. 60-90). Princeton: Princeton University Press.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

- Iramaneerat, C., Smith, Jr., E. V., & Smith, R. M. (2008). An introduction to Rasch measurement. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp.50-70). Los Angeles: Sage.
- Johanson, G. & Alsmadi, A. (2002). Differential person functioning. *Educational and Psychological Measurement*, 62(3), 435-443.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport: American Council on Education and Praeger.
- Kaput, J. J. (1987). Representation systems and mathematics. In C. Janvier (Ed.), *Problems of Representation in the Teaching and Learning of Mathematics* (pp. 19-26). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person fit statistics. *Applied Measurement in Education*, 16, 227-298.
- Klauer, K. C. (1995). The assessment of person fit. In G. Fischer & I. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 97-110). New York, NY: Springer-Verlag.
- Lamprianou, I. (2010). The practical application of optimal appropriateness measurement on empirical data using Rasch models. *Journal of Applied Measurement*, 11(4), 1-15.
- Lamprianou, I. (2013). The tendency of individuals to respond to high-stakes tests in idiosyncratic ways. *Journal of Applied Measurement*, 14(3), 299-317.
- Lamprianou, I., & Boyle, B. (2004). Accuracy of measurement in the context of

- mathematics national curriculum tests in England for ethnic minority pupils and pupils who speak English as an additional language. *Journal of Educational Measurement*, 41(3), 239-259.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269-290.
- Linacre, J. M. (2000). Item discrimination and infit mean-squares. *Rasch Measurement Transactions*, 14(2), 743.
- Linacre, J. M. (2011). Winsteps® (Version 3.73.0) [Computer software]. Beaverton, OR: Winsteps.com. Retrieved on February 20, 2011. Available from <http://www.winsteps.com/>
- Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36(8), 437-448.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635-694.
- Lord, F. & Novick, M. L. (1968). Measurement procedures and item-scoring formulas. In F. Lord & M. L. Novick (Eds.), *Statistical theories of mental test scores* (pp. 302-321). Reading, MA: Addison-Wesley Publishing Company.
- Lumsden, J. (1977). Person Reliability. *Applied Psychological Measurement*, 1, 477-482.
- Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, 31, 19-26. doi: 10.1111/j.2044-8317.1978.tb00568.x
- Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological*

Measurement, 4, 1-7.

- Mair, P., & Hatzinger, R. (2007a). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, 20(9), 1-20.
- Mair, P., & Hatzinger, R. (2007b). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science* 49(1), 26-43.
- Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, 8, 72-87.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25, 107-135.
- Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement*, 14(3), 283-298.
- Meijer, R. R., & van Krimpen-Stoop, E. M. (2010). Detecting person misfit in adaptive testing. In W. van der Linden & C. Glas (Eds.), *Elements of Adaptive Testing* (pp. 315-329). New York, NY: Springer.
- Messick, S. (1980). Test validity and the ethics of assessment. *American psychologist*, 35, 1012-1027.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
- Miller, M.D., Linn, R.L., & Gronlund, N.E. (2009). *Measurement and assessment in*

teaching, 10th Edition. Upper Saddle River, NJ: Prentice Hall.

Millman, J., Bishop, C.H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, 25, 707-726.

Molenaar, P. C. M. (2009). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives*, 2(4), 201-218.

Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*, 47, 355-366.

Mosier, C. I. (1941). Psychophysics and mental test theory II: The constant process. *Psychological Review*, 48, 235-249.

National Council of Teachers of Mathematics. (2001). *The roles of representation in school mathematics*. Reston, VA: NCTM.

Nering (1997). Distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement*, 21, 155-127.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement*, 22, 53-69.

Perkins A. & Engelhard, G., Jr. (2009). Crossing Person Response Functions. *Rasch Measurement Transactions*, 23(1), 1183-1184.

Perkins, A., Quaynor, L., & Engelhard, G. (2011). The influences of home language, gender, and social class on mathematics literacy in France, Germany, Hong Kong, and the United States. IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 4, 35-58.

Petridou, A., & Williams, J. (2007). Accounting for aberrant test response patterns using

- multilevel models. *Journal of Educational Measurement*, 44, 227-247.
- Petridou, A., & Williams, J. (2010). Accounting for unexpected test responses through examinees' and their teachers' explanations. *Assessment in Education: Principles, Policy & Practice*, 17, 357-382.
- R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.R-project.org>.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition. Chicago, IL: University of Chicago Press, 1980).
- Raudenbush, S.W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Yang, M. L., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, 9(1), 141-157.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.
- Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, 14, 127-137.
- Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research*, 35, 543-568.
- Rogers, H. J. (1999). Guessing in multiple-choice tests. In G. N. Masters and J. P. Keeves

- (Eds.). *Advances in Measurement in Educational Research and Assessment*. (pp. 23-42) Oxford, UK: Pergamon.
- Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Psychological Measurement*, 4, 159-183.
- Rogers, W. T., & Yang, P. Y. (1996). Test-wiseness: Its nature and application. *European Journal of Psychological Assessment*, 12, 247-259.
- Royal, K. D., & Hedgpeth, M. (2013). Investigating guessing strategies and their success rates on items of varying difficulty levels. *Rasch Measurement Transactions*, 27, 1407-1408.
- Rudner, L. M., Bracey, G., & Skaggs, G. (1996). The use of a person-fit statistic with one high-quality achievement test. *Applied Measurement in Education*, 9, 91-109.
- Seo, D.G., & Weiss, D. J. (2013). Iz Person-fit index to identify misfit students with achievement test data. *Educational and Psychological Measurement*, 73, 994-1016.
- Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in *Educational and Psychological Measurement*. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 91-111). Switzerland: Springer.
- Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, 63(2), 183-200.

- Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, 66, 191-208.
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to non-parametric item response theory*. Thousand Oaks, CA: Sage.
- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, 36, 477-481.
- Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19-37). Charlotte: Information Age Publishing.
- Sireci, S. G., & Forte, E. (2012). Informing the information age: How to communicate measurement concepts to educational policy makers. *Educational Measurement: Issues and Practice*, 31, 27-32.
- Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement*, 45, 433-444.
- Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, 46, 359-372.
- Smith, R. M. (1990). Theory and practice of fit. *Rasch Measurement Transactions*, 3, 78.
- Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51, 541-565.
- Smith, R. M. (1993). Guessing and the Rasch Model. *Rasch Measurement Transactions*, 6, 262-263.
- Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. Smith, Jr.,

- and R. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 73-92). Maple Grove, MN: JAM Press.
- Smith, R. M., & Plackner, C. (2010). The family approach to assessing fit in Rasch measurement. In M. L. Garner, G. Engelhard, Jr., W. P. Fisher, Jr. & M. Wilson (Eds.), *Advances in Rasch Measurement Volume 1* (pp. 64-85). Maple Grove, MN: JAM Press.
- Smith, R. M., Schumacker, R. E., & Busch, M. J. (1995). *Using item mean squares to evaluate fit to the Rasch model*. Paper presented at the 1995 Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Strandmark, N. L., & Linn, R. L. (1987). A generalized logistic item response model parameterizing test score inappropriateness. *Applied Psychological Measurement*, 11, 355-370.
- Swaminathan, H., Hambleton, R.K., & Rogers, H.J. (2007). Assessing the fit of item response theory models. In C. Rao, and S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (Vol. 26, pp. 683-718). Amsterdam, the Netherlands: Elsevier.
- Thurstone, L. L. (1926). Scoring of individual performance. *The Journal of Educational Psychology*, 17(7), 456-457.
- Tennant, A., & Pallant, J.F. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions*, 20(1), 1048-51.
- Trabin, T. E., & Weiss, D. J. (1979). *The person response curve: Fit of individuals to*

item characteristic curve models. (Research Report 79-7). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley Publishing Company.

Vale, D. C. & Weiss, D. J. (1975). *A study of computer-administered strataptive ability testing*. (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometrics Methods Program.

van den Noortgate, W., & Paek, I. (2004). Person regression models. In P. De Boeck, and M. Wilson (Eds.), *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach* (pp. 167-187). New York: Springer-Verlag.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. van der Linden and C. Glas (Eds.), *Elements of Adaptive Testing* (pp. 3-30). New York, NY: Springer.

van Krimpen-Stoop, E. M., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, 23, 327-345.

van Krimpen-Stoop, E. M., & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201-219). Dordrecht, The Netherlands: Kluwer Academic Publishers.

- Velleman, P. F., & Hoaglin, D.C. (1981). Smoothing data. In P. Velleman & D. Hoaglin (Eds.), *Applications, basics, and computing of exploratory data analysis* (pp. 159-199). Boston, MA: Duxbury Press.
- Walker, A. A., & Engelhard, G., Jr. (2014). Game-based assessments: A promising way to create idiographic perspectives. *Measurement: Interdisciplinary Research and Perspectives*, 12, 57-61.
- Walker, A. A., & Engelhard, G., Jr. (2015). Exploring person fit with an approach based on multilevel logistic regression. *Applied Measurement in Education*, doi: 10.1080/08957347.2015.1062767.
- Waller, M. I. (1976). *Estimating parameters in the Rasch model: Removing the effects of random guessing*. ETS-EB-76-8, Educational Testing Service, Princeton, New Jersey.
- Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement*, 13, 233-243.
- Wan, L., & Wu, B. (2008). *Person-fit of English Language Learners (ELL) in K-12 High-Stakes Assessments*. Retrieved from Pearson website: http://images.pearsonclinical.com/images/tmrs/tmrs_rg/PersonfitofELLinK12HighStakesAssessments.pdf
- Wang, L., Pan, W., & Bai, H. (2008). *Detection power of multilevel latent-trait differential person functioning: A Monte Carlo comparison with conventional person misfit statistics*. Paper presented at the 2008 International Meeting of Psychometric Society, Durham, New Hampshire.
- Weiss, D. J. (1973). *The stratified adaptive computerized ability test*. (Research Report

- 73-3). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometrics Methods Program.
- Wood, R. (1976). Inhibiting blind guessing: The effect of instructions. *Journal of Educational Measurement*, 13, 297-307.
- Woods, C. M. (2008). Monte Carlo evaluation of two-level logistic regression for assessing person fit. *Multivariate Behavioral Research*, 43, 50-76.
- Woods, C.M., Oltmans, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment*, 20(2), 159-168.
- Wright, B. D. (1968). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service.
- Wright, B. D. (1992). Dichotomous Model vs Birnbaum 3-PL Three-Parameter Logistic Model. *Rasch Measurement Transactions*, 5(4), 178.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.
- Wright B. D., Mead R. J., & Ludlow L. H. (1980). *KIDMAP: Person-by-Item Interaction Mapping*. MESA Memorandum #29. Chicago, IL: MESA Press.
- Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23-48.
- Wright, B. D., & Stone, M. H. (1979). *Best Test Design: Rasch Measurement*. Chicago, IL: MESA Press.
- Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, 9(4),

472.

- Wu, M. & Adams, F. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, 14(4), 339-355.
- Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, 31(2), 21-26.
- Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education*, 22, 359-375.
- Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.) *Psychometrics (Handbook of statistics, Vol. 26, pp.45-79)*. Amsterdam: Elsevier.
- Zumbo, B. D. (2009). Validity as a contextualized and pragmatic explanation, and its implications for validation practice. In R.W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp.65-82). Charlotte: IAP-Information Age Publishing.
- Zumbo, B. D., & Chan, E. K. H. (2014). Setting the stage for validity and validation in social, behavioral, and health sciences: Trends in validation practices. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 3-8). Switzerland: Springer.

Table 1. *Summary of Person Response Functions Research (1941 to 2014)*

Year(s)	Author(s)	Theoretical Alignment and Argument	Application Area
1941	Mosier	Psychophysics - score estimation; PRF used to illustrate how psychophysics methods can apply to obtaining “better” scores for test takers.	Educational Testing
1953	Carroll & Schohan	Forerunner to Item Response Theory; score estimation and interpretation; argues by using PRF stakeholders can obtain “better” scores for test takers where one can gauge level of “mastery” from a score not just “relative” standing in a group.	Educational Testing
1973, 1975, 1979	Weiss and colleagues: Weiss; Weiss & Vale; Trabin & Weiss	Item Response Theory model-data fit; argues if persons don’t fit the model, then the obtained measurement is not precise for them; items should be targeted to individuals (adaptive testing) to get the best estimates of ability. PRF can show alignment between model-expected and observed responses.	Educational Testing
1977, 1978	Lumsden	Item Response Theory test score interpretation; argues that more than a sufficient statistic (i.e., total score or weighted scale score) to evaluate student performance and that PRF can assist in interpretation and post-test instruction.	Educational Testing
1981	Brunk	Alternative to person parameter estimation; no argument for interpretation	Statistical Science
1987	Strandmark & Linn	Item Response Theory model-data fit; proposes a new generalized model that includes three person parameters (reliability-person slope and propensity to guess)	Educational Testing

Year(s)	Author(s)	Theoretical Alignment and Argument	Application Area
1990, 1991	Carroll and colleagues: Carroll; Carroll, Meade, & Johnson	Item Response Theory alternative to parameter estimation; argues that by using PRF, item and person parameters can be estimated without iteration and on small samples; PRF to check unidimensionality assumption	Educational Testing
2000	Reise	Item Response Theory model-data fit; argues that parameters of PRF can be modeled using multilevel logistic regression; model-data fit can be examined and person misfit explained and plotted	Personality Testing
2001,2004, 2005	Meijer and colleagues: Sijstma & Meijer; Emons, Sijstma, & Meijer	Item Response Theory & Non-Parametric Item Response Theory model-data fit and test score interpretation; argues that a multi-stage approach to detecting and evaluating person misfit is necessary; introduces several methods for IRT and N-IRT PRF.	Educational and Psychological Testing
2011, 2013	Engelhard and colleagues: Engelhard; Perkins, Quaynor & Engelhard;	Item Response Theory model-data fit and test score interpretation; argues that non-parametric PRF can convey information about student response patterns; argues that Differential Group Functioning may be illustrated using PRF	Educational Testing
2007, 2014	Ferrando	Factor analytic procedures and Item Response Theory “variable-theta” models: model-data fit and interpretation; argues that intra-test variation may not be person misfit, but a relevant characteristic of person behavior; introduces an approach to plotting PRF (based on the approach by Emons et. al, 2005)	Personality Testing

Table 2. *Parameter Estimates for the Multilevel Logistic Models*

	Model I	Model II	Model III
Reliability			
Person Achievement, β_{0jREL}	0.870	0.006	NA
Item Difficulty, β_{1jREL}	0.207	0.104	0.086
Coefficients			
Grand Mean Intercept, γ_{00}	0.818 (0.008)*	0.005 (0.003)	0.003 (0.003)
Person Achievement, γ_{01}	NA	1.010 (0.003)*	1.012 (0.003)*
Grand Mean Slope, γ_{10}	-0.991 (0.003)*	-1.027 (0.003)*	-1.026 (0.003)*
Variiances and Covariance			
Average Person Achievement, τ_{00}	1.247 (1.117)*	0.001 (0.031)	NA
Average Item Difficulty, τ_{11}	0.030 (0.172)*	0.014 (0.119)*	0.009 (0.096)*
Covariance, τ_{01}	-0.077	-0.004	NA

Note. Standard errors or standard deviations in parenthesis, where applicable.

NA-Not applicable for the model.

* - significantly different from 0, $p < .001$

Table 3. *Summary of Findings from the Anchored Rasch Calibration*

	Persons (N=31)	Items (N=63)
Measures		
<i>M</i>	-0.84	0.00
<i>SD</i>	0.47	0.99
<i>Count</i>	31	63
Infit		
<i>M</i>	1.22	1.33
<i>SD</i>	0.12	0.82
Outfit		
<i>M</i>	1.37	1.37
<i>SD</i>	0.21	0.83
Reliability of Separation	0.49	0.66

Note. Items were anchored to their historic difficulty parameters obtained from a well-targeted test-taker population.

Table 4. *Mean Person Fit Statistic Values at the 95th Percentile*

Achievement Cluster (θ)	Mean Value at 95th Percentile		
	Outfit	Infit	Bfit
-2	1.124	1.046	2.955
-1	1.205	1.052	3.048
0	1.144	1.050	2.957
1	1.201	1.050	3.050
2	1.030	1.047	3.024
	Misfit Threshold Values		
	Outfit	Infit	Bfit
	1.166	1.050	2.997

Note. The values reported represent the mean value at 95th percentile over 5 simulated datasets.

The misfit threshold values are the weighted average of the mean values at the 95th percentile. Individual person fit values greater than the thresholds indicated person misfit to the model.

Table 5. *Psychometric Information for 25 Examinees*

ID	Misfit Triggered	True Achievement Level	Estimated Achievement Level	Standard Error of Measure	Included in Figures?	Fit Statistic Values Outfit, Infit, Bfit	Absolute Bias
318	Outfit, Infit	1.73	2.24	0.38	No	2.59, 1.11, 1.85	0.51
322		1.04	1.21	0.32	No	0.87, 0.90, 2.25	0.17
648		-1.85	-1.33	0.32	No	0.84, 0.87, 1.41	0.52
699	All 3	1.36	1.61	0.36	No	1.24, 1.12, 3.16	0.25
750	Outfit	-0.62	-1.31	0.33	Figure 11	1.59, 1.03, 1.77	0.69
832	Infit	1.15	1.05	0.34	Figure 12	1.14, 1.13, 0.53	0.10
938		-0.08	0.41	0.33	Figure 11	0.88, 0.91, 0.20	0.50
1158		-1.11	-1.77	0.33	No	0.89, 0.93, 0.48	0.66
1254		0.50	0.48	0.33	Figure 11	1.04, 1.02, 1.99	0.02
1322	Bfit	1.20	1.92	0.34	No	1.00, 1.00, 7.95	0.72
1355		-1.28	-1.99	0.33	No	0.87, 0.92, 0.74	0.71
1431	Bfit	0.28	0.30	0.32	Figure 13	0.87, 0.89, 8.30	0.02
1487	Infit	0.24	0.71	0.36	No	1.17, 1.10, 0.61	0.47
1935		-0.17	-0.05	0.34	No	0.95, 0.95, 0.30	0.12
2227	Bfit	-1.05	-1.48	0.33	Figure 13	0.94, 0.97, 8.30	0.43
2554		-0.02	0.26	0.33	No	0.88, 0.91, 1.94	0.28
2566		0.66	0.63	0.34	No	0.92, 0.95, 1.16	0.03
2689		1.51	1.83	0.33	Figure 12	0.92, 0.95, 1.40	0.32
3638	Infit	-1.31	-1.78	0.35	Figure 12	1.08, 1.08, 0.77	0.47
3650	Outfit	-2.04	-1.31	0.33	Figure 11	1.74, 0.97, 0.22	0.73
4327	Infit	-0.47	-1.06	0.35	Figure 12	1.05, 1.07, 0.89	0.59
4403		2.18	2.04	0.34	Figure 12	0.93, 0.98, 0.39	0.14
4417	Outfit	1.57	1.24	0.34	Figure 11	1.52, 1.04, 0.14	0.33
4556	Bfit	-0.85	-1.53	0.32	Figure 13	0.87, 0.90, 8.64	0.68
4994	Infit	1.11	0.35	0.36	Figure 12	1.14, 1.09, 0.84	0.76

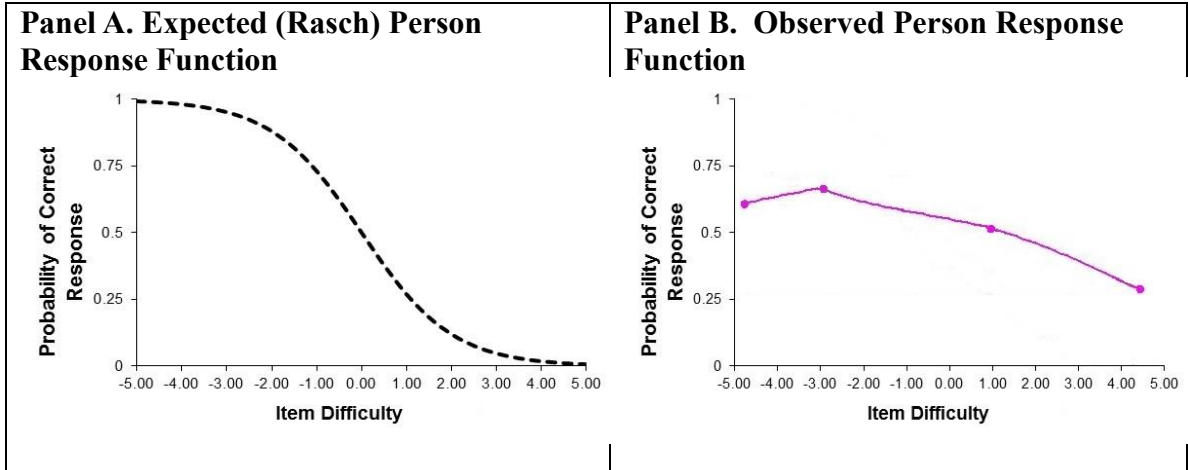


Figure 1. *Example Expected and Observed Person Response Function.*

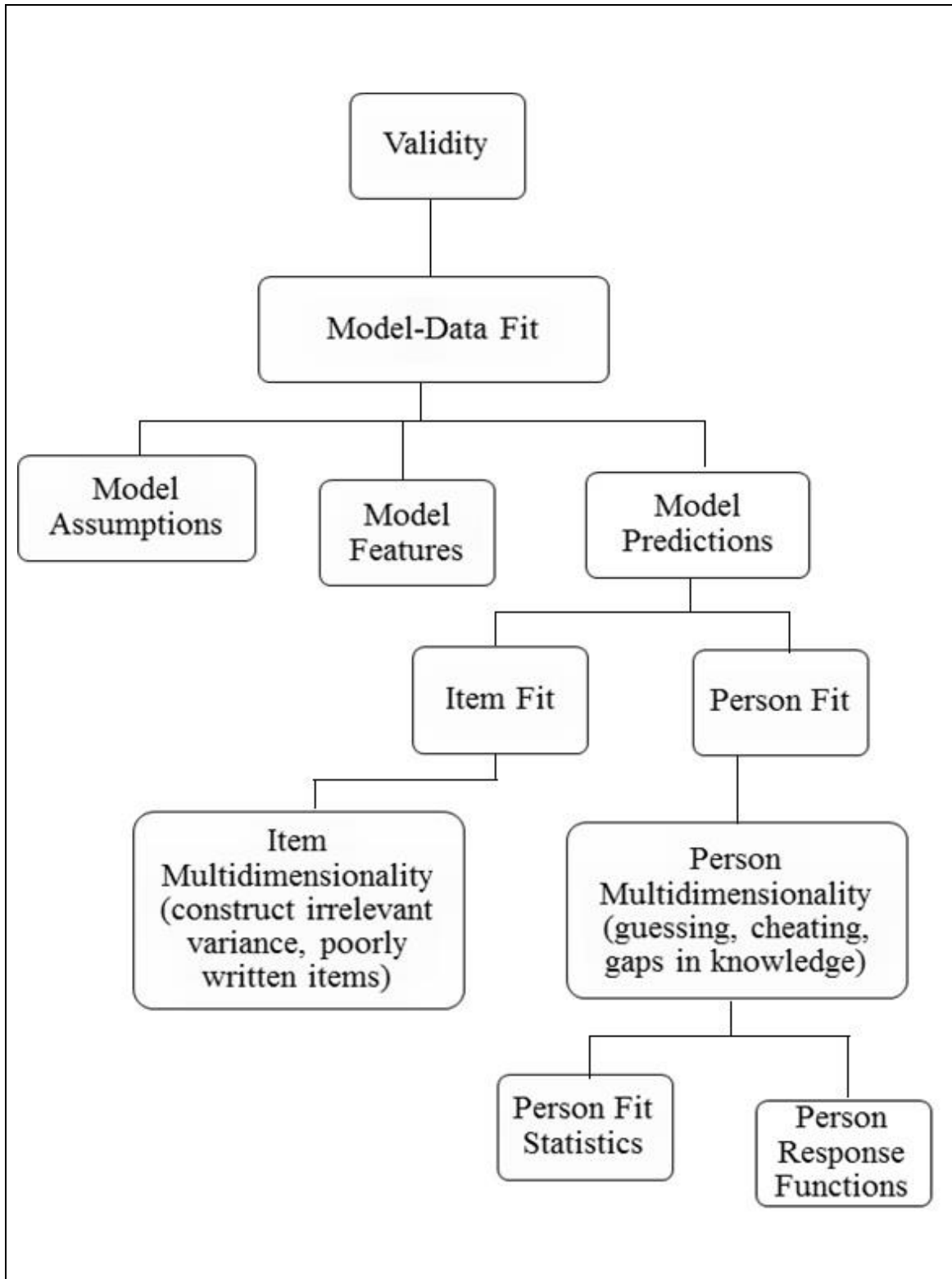


Figure 2. *Conceptual Outline.*

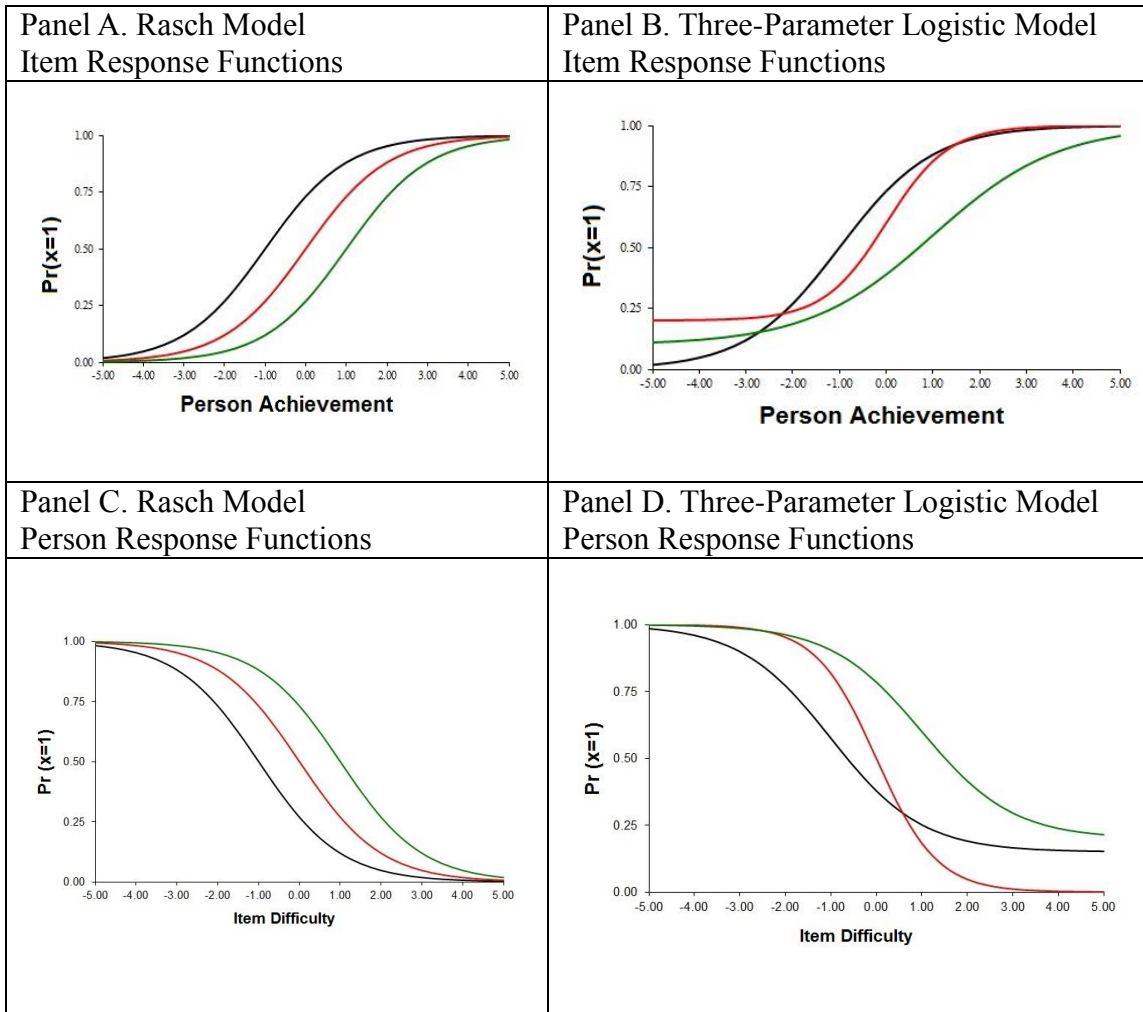


Figure 3. *Item and Person Response Functions for the Rasch and 3-PL.*¹²

¹² Note. In the Rasch model, the items are expected to differ only by difficulty and the persons are expected to differ only by achievement level. This single parameter is shown by the non-crossing item and person response functions. In the 3-PL model, the items are expected to differ also by how quickly they distinguish between persons of varying levels of achievement and how they may be correctly answered by guessing. Persons are expected to differ also by how quickly they distinguish between items of varying levels of difficulty and on their propensities to guess at items they don't know. These additional parameters are shown by the crossing item and person response functions (slope values or discrimination) and different lower asymptotes (pseudo-guessing). For this study, which uses the Rasch model, significant discrepancies in person or item discrimination or pseudo-guessing is considered misfit to the model.

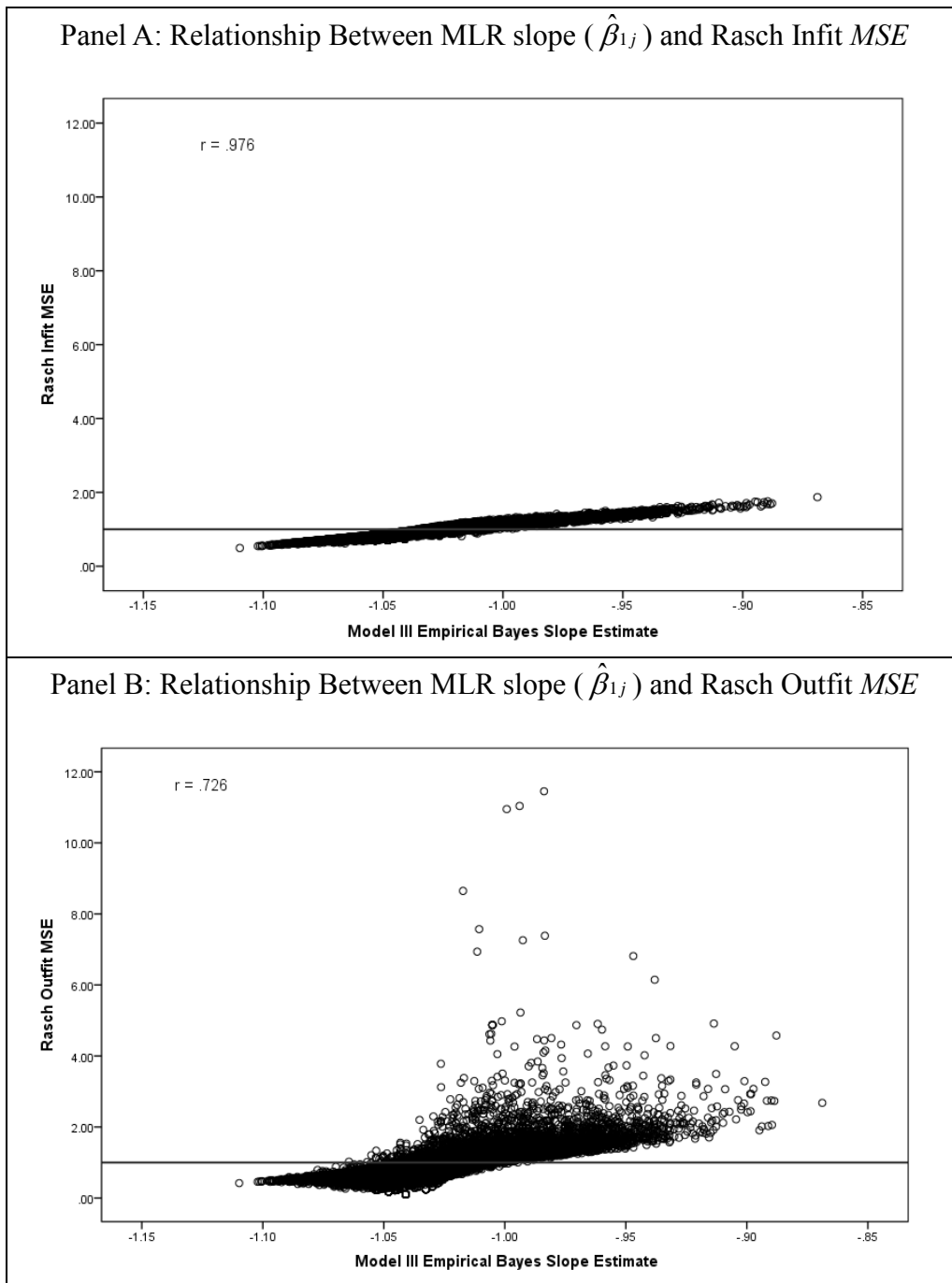


Figure 4. *Rasch Infit and Outfit Fit by MLR Estimates.*¹³

¹³ Note. The reference line at one on the y-axis represents the Rasch *MSE* value that is expected when good model-data fit is obtained. Approximately 77% of Infit *MSE* values and 45% of Outfit *MSE* values fall within the range of 0.8 and 1.2, which indicates reasonable model-data fit (Wright & Linacre, 1994).

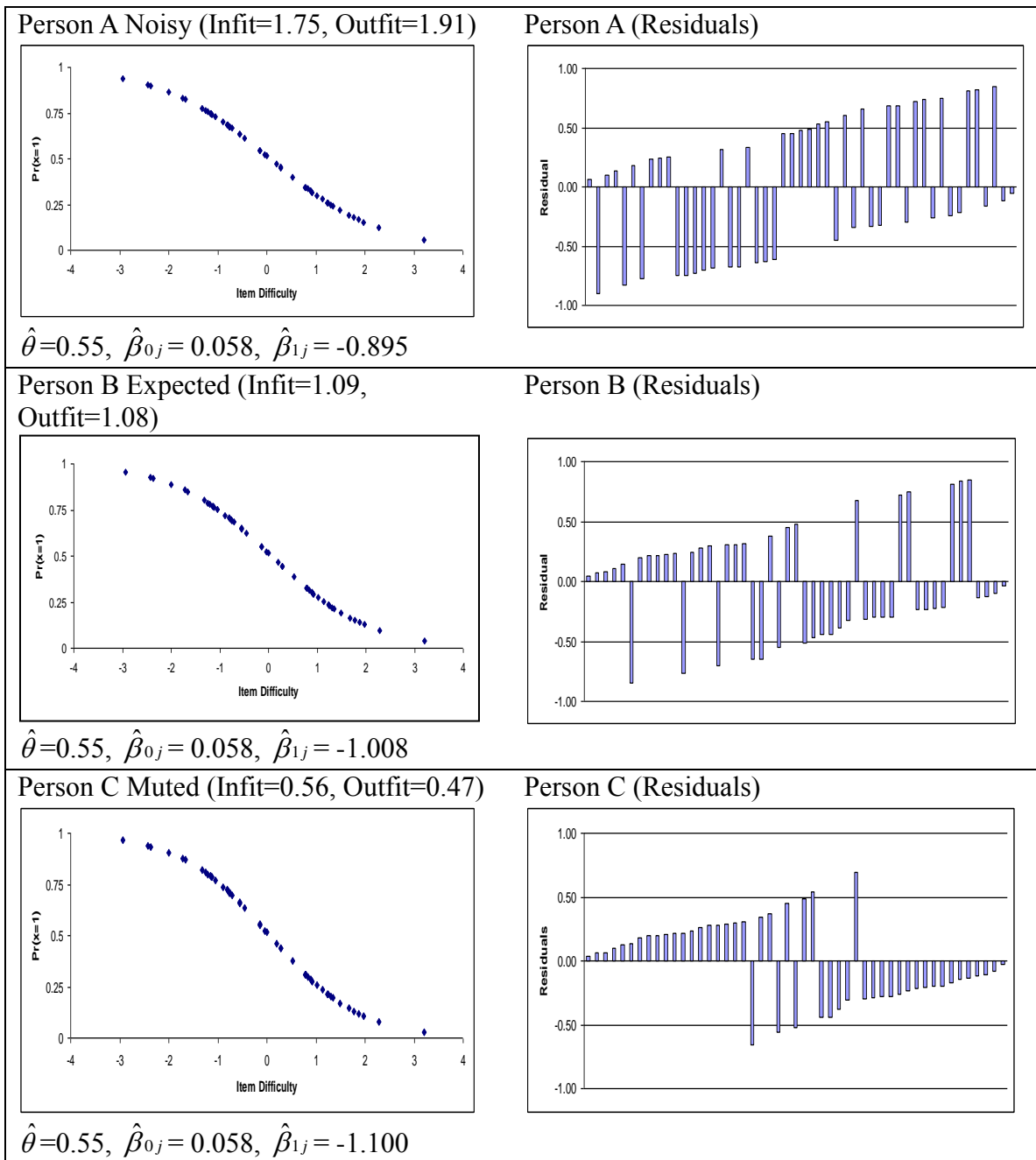


Figure 5. *Estimated person response functions and residuals.*

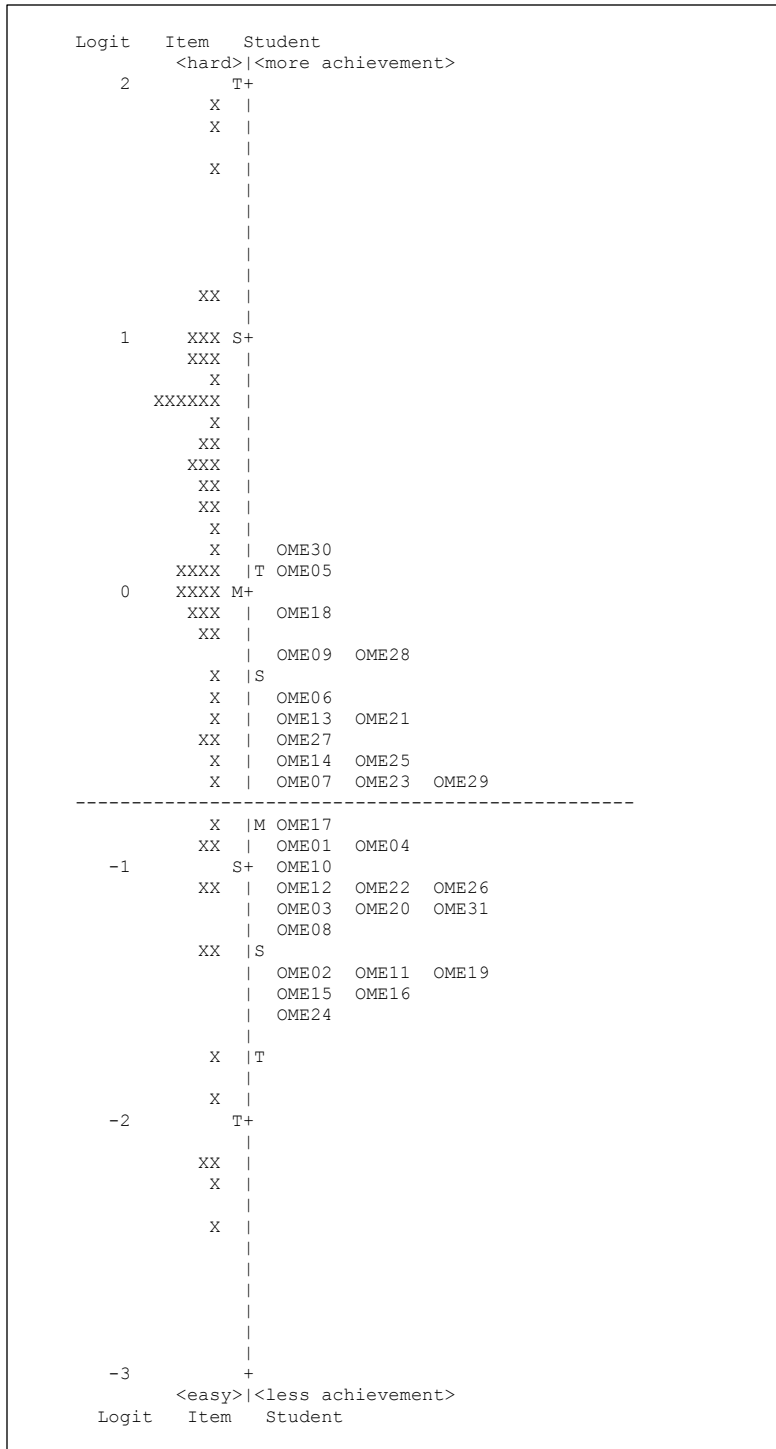


Figure 6. Variable Map for the Anchored Rasch Analysis.¹⁴

¹⁴ Note. A dashed reference line is drawn at -0.74 logits to represent the expected location for chance-level performance on this test.

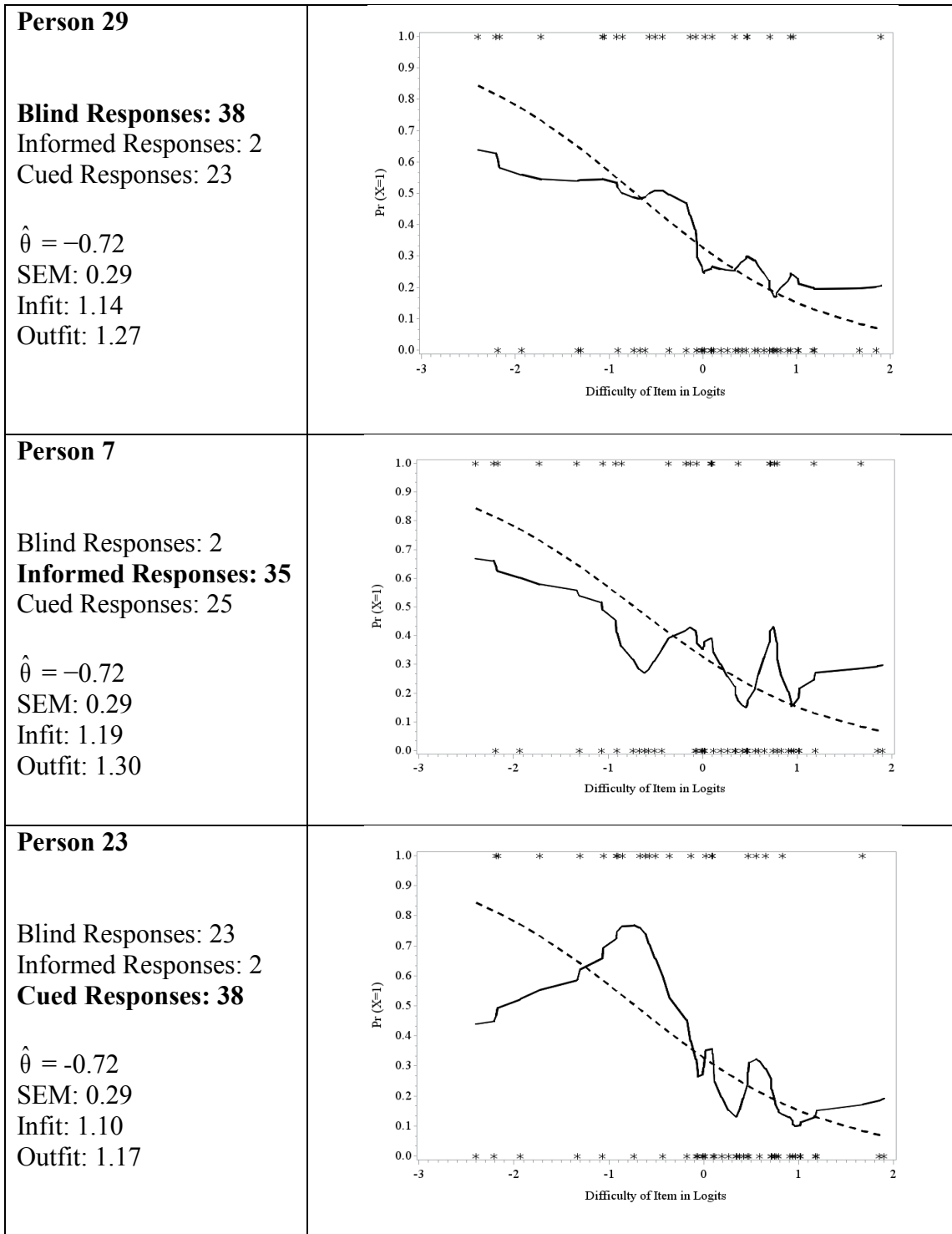


Figure 7. Fitting person response functions.

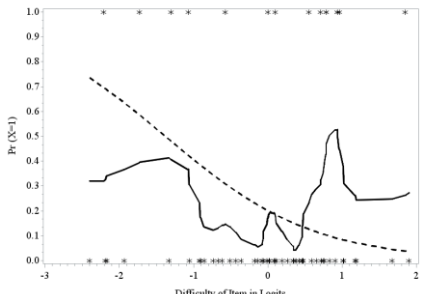
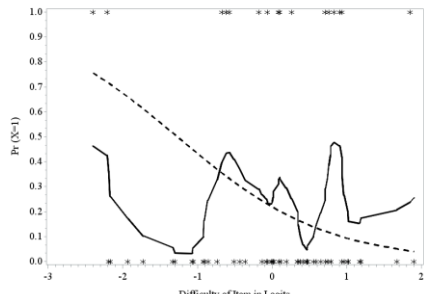
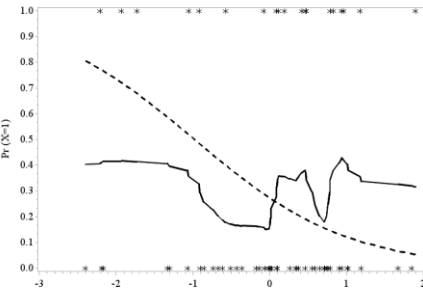
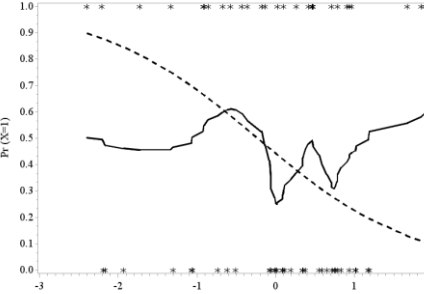
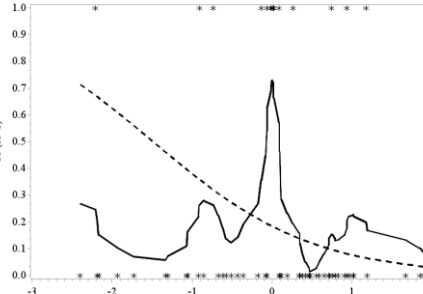
<p>Person 2 Blind Responses: 44 Informed Responses: 13 Cued Responses: 6</p> <p>$\hat{\theta} = -1.38$ SEM: 0.33 Infit: 1.41 Outfit: 1.95</p>		<p>Person 8 Blind Responses: 58 Informed Responses: 3 Cued Responses: 2</p> <p>$\hat{\theta} = -1.27$ SEM: 0.32 Infit: 1.41 Outfit: 1.76</p>	
<p>Person 10 Blind Responses: 62 Informed Responses: 1 Cued Responses: 0</p> <p>$\hat{\theta} = -0.98$ SEM: 0.30 Infit: 1.42 Outfit: 1.73</p>		<p>Person 9 Blind Responses: 57 Informed Responses: 6 Cued Responses: 0</p> <p>$\hat{\theta} = -0.23$ SEM: 0.28 Infit: 1.38 Outfit: 1.64</p>	
<p>Person 16 Blind Responses: 59 Informed Responses: 4 Cued Responses: 0</p> <p>$\hat{\theta} = -1.49$ SEM: 0.33 Infit: 1.43 Outfit: 1.52</p>		<p>Intentionally Blank</p>	

Figure 8. Misfitting person response functions, blind guessing

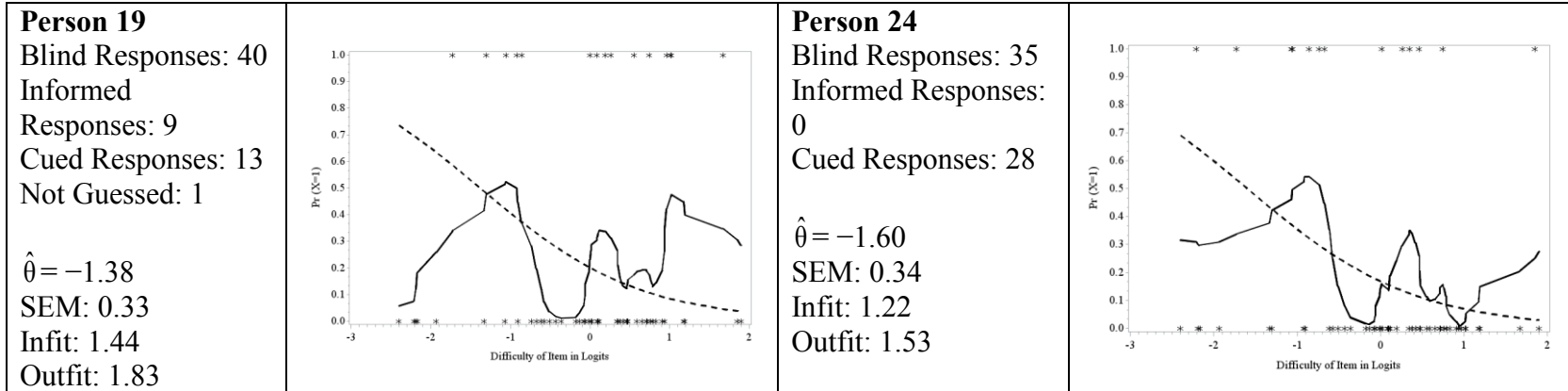


Figure 9. Misfitting person response functions, other guessing.

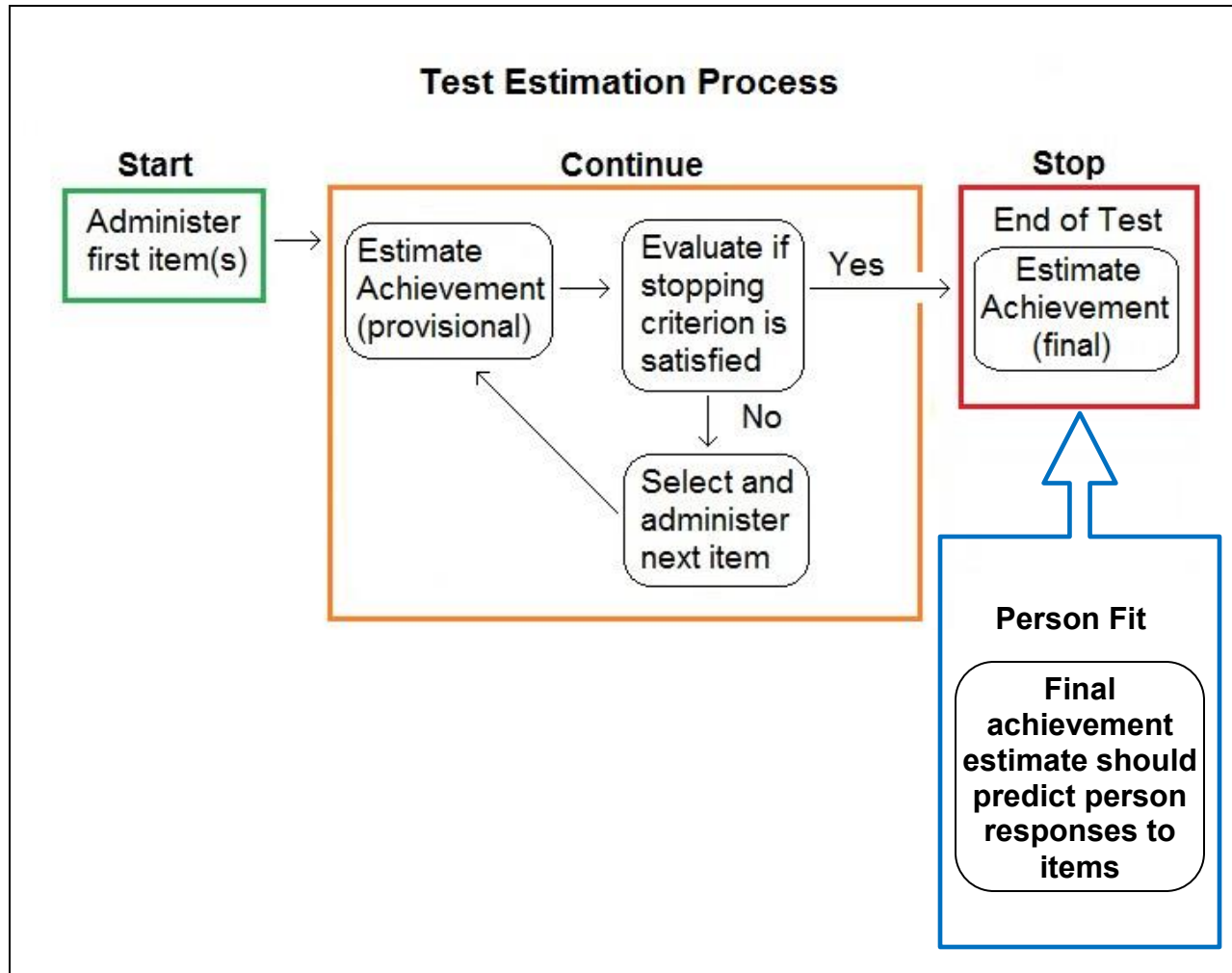


Figure 10. *Conceptual framework of person fit in CAT.*

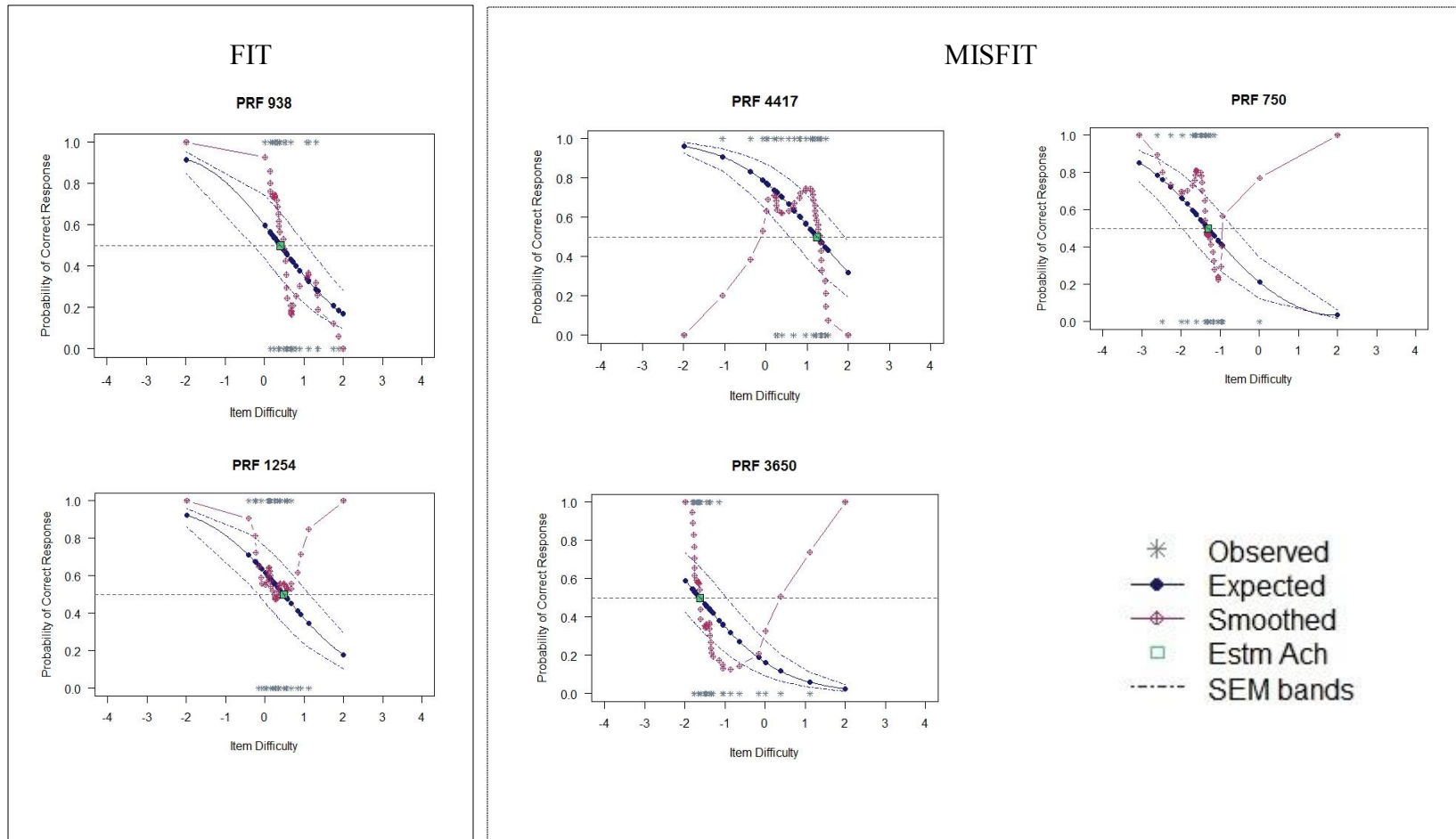


Figure 11. *Person response functions illustrating misfit by Outfit.*¹⁵

¹⁵ Note. Item difficulty is represented on the x-axis and the probability of giving the correct response is represented on the y-axis. The reference line is included at $Pr(x=1)=0.50$, the location of the achievement estimate in the Rasch model. The first box shows 2 response patterns that fit the model; the second box shows 3 response patterns that do not fit the model.

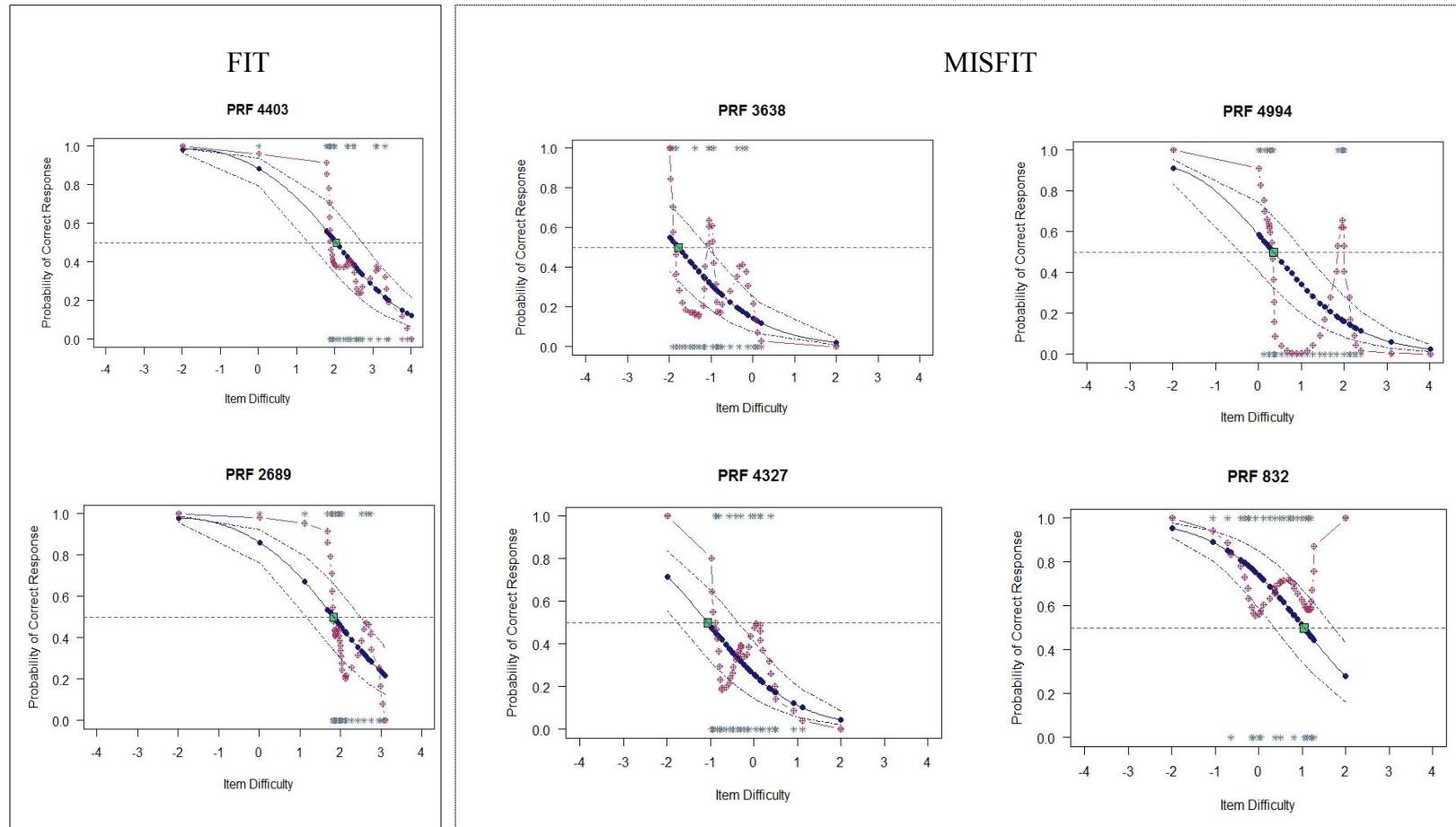


Figure 12. *Person response functions illustrating misfit by Infit.*¹⁶

¹⁶ Note. Item difficulty is represented on the x-axis and the probability of giving the correct response is represented on the y-axis. The reference line is included at $Pr(x=1)=0.50$, the location of the achievement estimate in the Rasch model. The first box shows 2 response patterns that fit the model; the second box shows 4 response patterns that do not fit the model.

MISFIT

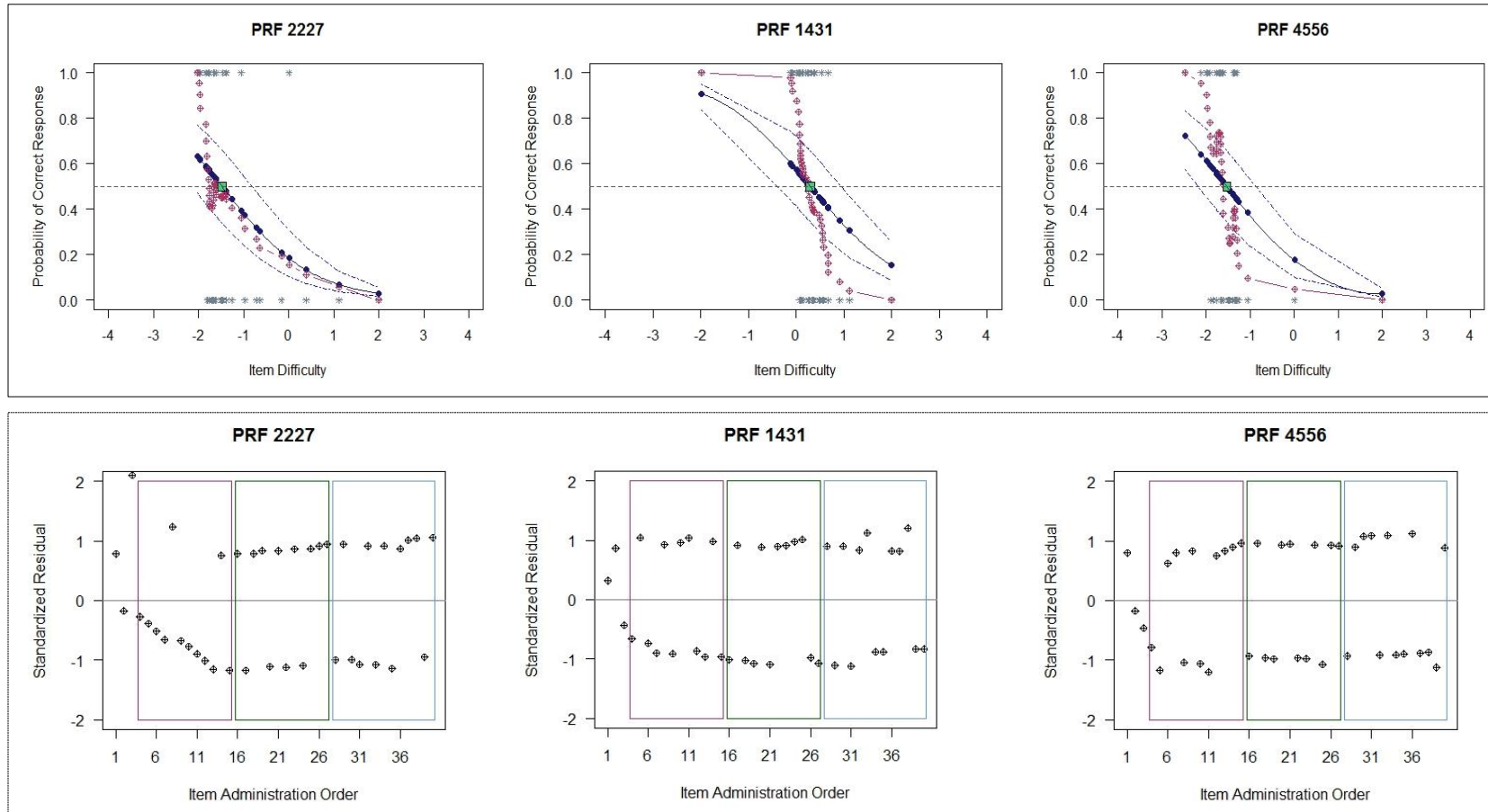



Figure 13. *Person response functions illustrating misfit by Bfit.*¹⁷

¹⁷ Note. In the top row, item difficulty is represented on the x-axis and the probability of giving the correct response is represented on the y-axis. The reference line is included at $\Pr(x=1)=0.50$, the location of the achievement estimate in the Rasch model. In the bottom row, an alternate view of misfit is illustrated. The x-axis represents item administration order and the y-axis represents the standardized residual.

Appendix A.
IRB Determination Letter, Application One

	EMORY UNIVERSITY	Institutional Review Board
<p>To: Angela Walker, Principal Investigator RE: Exemption of Human Subjects Research, IRB00059827</p>		
<p><i>Explanatory Person-Fit Analyses with Statistical and Graphical Approaches Based on Multilevel Logistic Regression</i></p>		
<p>Dear Principal Investigator:</p>		
<p>Thank you for submitting an application to the Emory IRB for the above-referenced project. Based on the information you have provided, we have determined on August 21, 2012 that although it is human subjects research, it is exempt from further IRB review and approval.</p>		
<p>This determination is good indefinitely unless substantive revisions to the study design (e.g., population or type of data to be obtained) occur which alter our analysis. Please consult the Emory IRB for clarification in case of such a change. Exempt projects do not require continuing renewal applications.</p>		
<p>This project meets the criteria for exemption under 45 CFR 46.101(b)(2). Specifically, you will be using SATW results to determine a new way to evaluate whether or not standardized test scores on the SATW are good indicators of student achievement in writing.</p>		
<p>Please note that the Belmont Report principles apply to this research: respect for persons, beneficence, and justice. You should use the informed consent materials reviewed by the IRB unless a waiver of consent was granted. Similarly, if HIPAA applies to this project, you should use the HIPAA patient authorization and revocation materials reviewed by the IRB unless a waiver was granted. CITI certification is required of all personnel conducting this research. Unanticipated problems involving risk to subjects or others or violations of the HIPAA Privacy Rule must be reported promptly to the Emory IRB and the sponsoring agency (if any).</p>		
<p>In future correspondence about this matter, please refer to the study ID shown above. Thank you.</p>		
<p>Sincerely, Rebecca Rousselle, CIP Assistant Director</p>		
<p><i>This letter has been digitally signed</i></p>		
<p>CC: Engelhard, George Educational Studies</p>		

Appendix B.
IRB Determination Letter, Application Two



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

OFFICE OF HUMAN RESEARCH ETHICS

Medical School Building 52
Mason Farm Road
CB #7097
Chapel Hill, NC 27599-7097
(919) 966-3113
Web site: ohre.unc.edu
Federalwide Assurance (FWA) #4801

To: Kenneth Royal
Medical Education

From: Office of Human Research Ethics

Date: 2/25/2013

RE: Notice of IRB Exemption

Exemption Category: 1.Educational setting,2.Survey, interview, public observation

Study #: 13-1431

Study Title: Investigating the impact of good test-taking skills on a medical school exam

This submission has been reviewed by the Office of Human Research Ethics and was determined to be exempt from further review according to the regulatory category cited above under 45 CFR 46.101(b).

Study Description:

Purpose: To investigate non-medical students performance on a medical school exam; Are good test-taking skills and minimal medical content knowledge sufficient to pass?

Participants: A small sample of UNC professional staff volunteers from the Offices of Medical Education (OME)

Procedures (methods): A medical school exam will be arbitrarily chosen and OME staff will attempt to pass the exam. OME staff performance will be compared to medical students' performance. It is anticipated that OME staff scores will be comparable to odds of random guessing (approximately 20-25% correct).

Investigator's Responsibilities:

If your study protocol changes in such a way that exempt status would no longer apply, you should contact the above IRB before making the changes. The IRB will maintain records for this study for 3 years, at which time you will be contacted about the status of the study.

Researchers are reminded that additional approvals may be needed from relevant "gatekeepers" to access subjects (e.g., principals, facility directors, healthcare system).

Appendix C.

IRB Determination Letter, Application Three

Because this study used simulated data, no IRB review was necessary.