**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:

| | |
|---|---|
| Aminah Farita Perkins | Date |

Differential Person Functioning

By

Aminah F. Perkins
Doctor of Philosophy
Division of Educational Studies

---

George Engelhard Jr., Ph.D.
Advisor

---

Yuk Fai Cheong, Ph.D.
Committee Member

---

Robert Jensen, Ph.D.
Committee Member

Accepted:

---

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

---

Date

# Differential Person Functioning

By

Aminah F. Perkins
B.S., Spelman College, 2004
M.A., University of Georgia, 2008

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
2013

# Abstract

## Differential Person Functioning

By Aminah F. Perkins

The accuracy and meaningfulness of test scores is a crucial issue in educational settings marked by high-stakes assessments within No Child Left Behind and Race to the Top. Differential person functioning (DPF) is presented in this study using a Rasch measurement framework as a means for assessing the accuracy and validity of scores on educational assessments. The purpose of this study is to further our current understanding of DPF as not only a threat to test score validity, but as a way to examine individual student response behaviors. Erasure analyses and multilevel modeling are used as the methods to identify and assess DPF across various contexts. The following questions are used to guide the research:

(1) What is differential person functioning?
(2) How do the methods for assessing differential person functioning differ across contexts?
(3) To what extent does differential person functioning contribute to our understanding of person fit across contexts?

The first question is answered through an extensive review of literature on the various components of DPF: person measurement, person response functions, person fit indices, and response behaviors. Guiding questions (2) and (3) are explored using data from a high-stakes third grade statewide assessment of mathematics and reading achievement. These questions are explored using two case studies, each replicated within two content areas (mathematics and reading) yielding a total of four contexts that are explored. The first case study investigates the relationship between wrong-to-right erasures, person fit indices, and school-level mathematics and reading achievement using the Many Facets Rasch model (MFRM) and a pre/post erasure design. The second case study uses hierarchical generalized linear modeling (HGLM) to examine student and school factors that may be associated with the aberrant responses of students that include proficiency levels, economic status, gender, and erasure behavior.

The dissertation sheds light on the importance of evaluating DPF when considering the validity evidence for an assessment. Additionally, MFRM and HGLM, yielded valuable information for researchers to begin to consider systematic routine analyses of DPF for high stakes assessments.

Differential Person Functioning

By

Aminah F. Perkins
B.S., Spelman College, 2004
M.A., University of Georgia, 2008

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
2013

**ACKNOWLEDGEMENTS**

*Children are the reward of life.*

*African proverb*

This dissertation is dedicated to the subjects of my research, the countless children in this country who depend on us, researchers, teachers, educators, policy makers, parents, and caregivers to provide them with the best education possible.

*We must remember that intelligence is not enough. Intelligence plus character--that is the goal of true education. The complete education gives one not only power of concentration, but worthy objectives upon which to concentrate.*

*Martin Luther King, Jr.*

I would like to first acknowledge my advisor, Dr. George Engelhard, Jr. I thank you for your endless support and encouragement. I thank you for being more than an advisor, for being a mentor and a role model in academia. To my committee members, Dr. Yuk Fai Cheong and Dr. Robert Jensen I am extremely appreciative of your advice and never ending willingness to offer support. I would also like to thank my cohort Nadia Behizadeh, Sheryl Croft, Brandi Hinnant-Crawford, Miyoshi Juergensen, Tiffany Pogue, and Zendre Sanders, for their support and encouragement on this journey that on many days seemed never ending. It was in the Division of Educational Studies that I found another home where I had the pleasure of learning from a dynamic faculty and profoundly intelligent colleagues. It is in this home that my knowledge of the world was expanded. I am honored to be among the many graduates of DES dedicated to the improvement of our educational system.

To my love, Worth Kamili Hayes, I humbly thank you for your encouragement and for enduring my stress and anxiety. Watching you on your journey was my best example of discipline and hard work. Thank you for giving up your work days so that I could have more time to focus on my dissertation. And to my son, Kamili Bakari, mommy thanks you for your laughter, smiles, and kisses. Know that what I do is for you and children like you.

*Live as if you were to die tomorrow. Learn as if you were to live forever.*

*Mahatma Gandhi*

My best example of a life-long learner has been my mother, Ruth Yvonne Perkins. My appreciation for all that you have been for me can never been quantified. I love you always. To my family, particularly, Fatima Hill, Eddie Hill, Felecia Bedford, Phillip Perkins, Nell Dotson, and Jameelah Carver, your love and encouragement on this journey mean more than you know. And to my extended family the Hayes' thank you for loving me like a daughter, sister and friend. I would also like to acknowledge the path laid for me by my ancestors, particularly my grandmother, Mary Dotson.

My friends and colleagues have been by my side holding me up when times were hard, providing countless edits to manuscripts and keeping me focused on the end goal. A special thank you to Jade Caines, Curtis Goings, Keisha Green, Charlotte Newman, Beryl Otumfuor, Michelle Purdy, Laura Quaynor, Ellana Stinson, Andrea Tullis, and Vincent Willis.

*Education is the most powerful weapon which you can use to change the world.*

*Nelson Mandela*

One of the greatest contributors to my development as a scholar, an activist, and a woman was my alma mater, Spelman College. I hope that my forthcoming accomplishments will one day grant me the honor of standing in the ranks of the many great women who came before me.

This research has also been supported by funding obtained from the American Educational Research Association Minority Dissertation Fellowship.

To those whose names are not listed, I apologize for the omission and please know that your contributions to my journey have not gone without notice. Most importantly, I thank and acknowledge the Creator of this world and all the worlds for providing me with limitless opportunities and the means to reach for and achieve my goals.

Lastly, I ask for the reader's forgiveness of all errors and limitations and I charge those who read this study to add to the research I and countless others before me have begun. The struggle continues…

**TABLE OF CONTENTS**

**List of Tables**

List of Figures

# CHAPTER ONE: INTRODUCTION

The current educational climate marked by federal programs, such as *No Child Left Behind* and *Race to the Top,* rely strongly on high-stakes testing. Despite debates regarding educational policies that place a deep reliance on high-stakes assessment, test developers are tasked with the role of ensuring that the information garnered from tests are accurate and fair. Test score accuracy (and in turn, inaccuracy) can affect the lives of students and educators in countless ways. Because the decisions based on tests can carry significant consequences, test developers are responsible for being aware of the many possible threats to validity that may occur when constructing and using assessments.

Messick (1989) described validity as "an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use" (p. 13). Data-to-model misfit occurs in the form of construct-irrelevant variance, a well-known threat to validity. Construct-irrelevant variance, defined more specifically as skills or characteristics of the examinee that are not intended to be assessed by the test (Ackerman, 1992), can exist in the form of differential item functioning (DIF) and differential person functioning (DPF). Differential item and person functioning exist as two possible threats to the validity of assessments. As Hambleton (1989) points out, "a poorly fitting model cannot yield invariant item- and ability- parameter estimates" (p. 172). DIF, a well known concept in the measurement literature is defined by Clauser and Mazor (1998) as the differing probabilities of success on an item between groups after they have been matched on a latent trait. The present study explores the concept of DPF, "an alternative to the usual DIF analysis" (Johanson & Alsmadi, 2002, p. 435). DPF is defined as unexpected differences between the observed and expected performance of persons on a set of items. This study serves as an exploration of DPF by examining the impact of DPF across various contexts.

A concept closely related to DPF is that of person fit. Person fit analysis is a psychometric approach used to assess the "believeability" of a person's individual response pattern on an assessment (Meijer, 1996; Smith, 1986). Person fit analysis can be used to provide numerical estimates of the degree to which individual response patterns are what would be expected given the measurement model used to assess the data. A statistically significant amount of person variation on an assessment can impact the validity of the assessment. Additionally, if DPF is present even for one individual, this could impact the validity of the assessment for that particular individual. In which case, a qualitative appraisal of the individual can prove useful in understanding the individual's unique interaction with the assessment.

If an assessment is free of a statistically significant amount of both DIF and DPF, then the latent variable measured by the assessment can be mapped onto a scale. This scale defines the latent variable under study, providing a description of what might be expected of people at different levels on the variable (Wilson, 2005). The existence of this theoretical latent variable should be supported empirically, and evaluated in terms of data-to-model fit. An assessment that does not meet the requirements of invariant measurement will not have good data-to-model fit.

Within item response theory (IRT), there exists a duality between person-invariant calibration of items (no DIF) and item-invariant measurement of persons (no DPF). As early as 1940, Mosier raised the idea of person and item invariance in the area of psychophysics. In particular, Mosier (1940) emphasized the necessity of taking "into account the variability of the individual” with respect to a set of items (p. 356). More specifically, Mosier recognized that the reliability of a set of items may vary from one individual to the next. This idea is at the heart of differential person functioning and person fit analysis.

## Theoretical Framework

The presence of differential person functioning (DPF) signifies an issue with the validity of the assessment for an individual. In the traditional interpretation of item response theory (IRT), the presence of DPF would be considered a threat to validity implying that the assessment is measuring a construct that was not intended to be measured by the assessment – construct *irrelevant* variance. However, one can argue that the person factors influencing DPF are indeed *relevant* factors for interpreting the meaning of the responses and scores of a given individual. For an assessment to measure the same construct in any population (invariant measurement) a certain set of core assumptions must hold. The requirements for invariant measurement as described by Engelhard (2013) are as follows:

*Item calibration:*

1. The calibration of the items must be independent of the particular persons used for calibration: *Person-invariant calibration of test items.*
2. Any person must have a better chance of success on an easy item than on a more difficult item: *Non-crossing item response functions.*

*Person measurement:*

3. The measurement of persons must be independent of the particular items that happen to be used for the measuring: *Item-invariant measurement of persons.*

4. A more able person must always have a better chance of success on an item than a less able person: *Non-crossing person response functions.*

*Variable map:*

5. Person and items must be located on a single underlying latent variable: *Unidimensionality*.

In particular, requirements (3) and (4) address issues related to differential person functioning and person fit analysis.

Rasch (1960/1980) identifies specific objectivity as a situation in which the relationship between two items is independent of the participants used for the comparison (invariant measurement). Wright (1967) supports this notion of "objective measurement" in the following statement:

> First, the calibration of measuring instruments must be independent of those objects that happen to be used for the calibration. Second, the measurement of objects must be independent of the instrument that happens to be used for the measuring. In practice, these conditions can only be approximated. But their approximation is what makes measurement objective (p. 87).

These properties are necessary conditions for the development of scales that meet the requirements of invariant measurement.

**Rasch Measurement Theory**

Rasch (1960/1980) measurement theory allows for the development of measures that adhere to the requirements for invariant measurement. Rasch measurement models enable the conception of measurement scales in the form of a ruler. Envisioning the ruler as a continuum on which a latent variable of interest lies there would exist more of the trait on one end and less of the trait at the other. Items can then be placed along this line at points corresponding to the amount of the trait that they require for endorsement. Individuals can also be placed on this line corresponding to the location at which they will endorse most of the items below their location on the line. In Rasch measurement, this construction is referred to as a variable map.

The relationship between persons and items can be modeled mathematically. Operating characteristic functions (OCFs) for dichotomous responses have been proposed by Rasch (1960/1980). The Rasch model for dichotomous responses can be written as

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

where, $\phi_{ni}$ represents the probability of endorsing an item given a person n with location $\theta_n$ on the latent variable, and item *i* with a difficulty (or location ) of $\delta_i$. In particular, this study focuses on the use of the Rasch model as the IRT model for analysis. However, it is important to note that Birnbaum (1968) also proposed OCFs for dichotomous responses in which the additional parameters of discrimination (ability of the item to differentiate between individuals at different locations on the latent variable) and pseudo-guessing (probability that a person with a low location on the latent variable will endorse an item by chance) are included. The Birnbaum model for dichotomous responses is

$$\phi_{ni} = c_i + (1 - c_i) \frac{\exp(\alpha_i (\theta_n - \delta_i))}{1 + \exp(\alpha_i (\theta_n - \delta_i))} \qquad [2]$$

where,

$\alpha_i$ = discrimination parameter for item i in the Birnbaum model, and

$c_i$ = lower asymptote of the function in the Birnbaum model often referred to as a pseudo-guessing parameter.

When the data fits the proposed IRT model, at higher person locations there exists a greater probability of endorsement of items. At lower person locations there exists a lower probability of endorsement of items. For example, we would expect that a high achieving student in pre-algebra would have a high probability of obtaining correct answers on pre-algebra items. While a student with a lower achievement level in pre-algebra would be expected to have a lower probability of obtaining correct answers on the same algebra items. A graphical representation of this relationship exists as an item response function (IRF), a monotonically increasing ogive.

If we select a particular person, such as Person A, then Equations 1 and 2 can be used to define person response functions (PRFs). The PRF utilizes the same mathematical model used for IRT models (Carroll, Meade, & Johnson, 1991). The Rasch PRF for Person A is

$$\phi_{Ai} = \frac{\exp(\theta_A - \delta_i)}{1 + \exp(\theta_A - \delta_i)} \qquad [3]$$

while the Birnbaum PRF is

$$\phi_{Ai} = c_A + (1 - c_A)\frac{\exp(\alpha_A(\theta_n - \delta_i))}{1 + \exp(\alpha_A(\theta_n - \delta_i))} \qquad [4]$$

It should be noted that $c_A$ is conceptually closer to a real "guessing" parameter in the Birnbaum PRFs, and that $\alpha_A$ represents person sensitivity or reliability to a particular subset of items. Carroll, Meade, and Johnson (1991) note that the only way in which the PRF differs from the IRF is that "the probabilities yielded by the equation are to be studied for a single individual (or group of individuals with similar values of $\theta$) as a function of different values of *b*, for different tasks" (p. 110).

**Fit Indices**

Fit indices can be utilized to evaluate item and individual fit to a given model. Researchers have suggested a multitude of fit statistics each with their own advantages and disadvantages (Karabatsos, 2003; Li & Olejnik, 1997; Reise 1990; Rudner 1983; Rudner, Bracey, & Skaggs, 1996; Sijtsma & Meijer, 2001; Smith, 1986). Given that the present study will employ the Rasch model, the traditional statistics of Outfit Mean Square (MNSQ) and Infit Mean Square (MNSQ) will be used to asses fit. Outfit MNSQ and Infit MNSQ provide a quantitative measure of the degree to which items or persons deviate from the expected model. To understand the calculation of Outfit MNSQ and Infit MNSQ we must first discuss the calculation of a residual. A response residual is a calculation of how far a person response ($x_{ni}$) deviates from an expected response ($E_{ni}$; Bond & Fox, 2007).

$$y_{ni} = x_{ni} - E_{ni} \qquad [5]$$

Outfit MNSQ and Infit MNSQ are used to quantify in one measure many person-item deviations. Outfit MNSQ is a measure of fit that is more sensitive to outliers and Infit MNSQ is a measure of fit more sensitive to inliers (Linacre, 2009). The Infit MNSQ for a person is the sum of the squared-standardized residuals, $Z^{2}_{ni}$, summed over the individual's response to all items. This variance is then averaged by dividing it by the number of items the individual responded to and is then weighted by the individuals variance ($W_{ni}$) to account for the impact of the outliers, resulting in an Infit measure as seen in Equation 6 (Bond & Fox, 2007; Petridou & Williams, 2007). For this reason, Infit is referred to as the information-weighted sum.

$$Infit = \frac{\sum Z_{ni}^{2} W_{ni}}{\sum W_{ni}} \qquad [6]$$

The Outfit MNSQ statistic is calculated similarly as seen in Equation 7. The difference lies in the fact that the residuals are not weighted.

$$Outfit = \frac{\sum Z_{ni}^{2}}{N} \qquad [7]$$

While many researchers have suggested other statistics as measures of person fit, Outfit and Infit were chosen for the current study given the reasons outlined above.

**Statement of the Problem**

While differential item functioning is a highly regarded concept familiar to most psychometricians (Zumbo, 1999), the current measurement literature addresses issues of person reliability or person variability less frequently. This study stresses the idea that data-to-model fit can be conceptualized in terms of both item response functions (IRFs) and person response functions. Perkins, Quaynor, and Engelhard (2011) and Engelhard (2009) suggest that researchers should begin to think more systematically about differential person functioning. It is important to recognize that items may function differently over different subgroups of persons. There is diversity among individuals who are often considered a homogeneous group such as women, English Language Learners, and African Americans. It is also important to recognize that persons may not function as intended in their interactions with subsets of test items. Guttman (1944) best describes the importance of assessing differential item and person functioning in the following statement:

> If a universe is scalable for one population but not for another population or forms a scale in a different manner, we cannot compare the two populations in degree and say that one is higher or lower on the average than another with respect to the universe (p. 1950).

This work seeks to evaluate the utility of differential person functioning and person fit in assessing individual performance.

## Purpose of the Study

The purpose of this dissertation is to examine the usefulness of differential person functioning as a method to assess invariant measurement. The study applies this method of analysis using two case studies replicated in the mathematics and reading content areas. Previous research potential biases related to variations in item difficulty for different students while also showing that the reliability of items vary from one individual to the next just as Mosier discussed as early as 1940 (Perkins, 2010). However, research illustrating how the assessment of DPF might differ based on context is lacking from current research. Additionally, while many simulation studies have been proposed for person fit research (Armstrong, Stoumbous, Kung, & Shi, 2007; Karabatsos, 2003; Levine & Rubin, 1979; Li & Olejnik, 1997; Meijer, Muijtjens, & van der Vleuten, 1996; Rudner 1983; Wright, 1977), there exists a lack of empirical research in the area. The present study seeks to bridge some of the gaps in the current measurement literature. This study can be used to demonstrate the use of person fit research, yet one must recognize that "whether person-fit statistics can help a research in practice depends on the context in which research takes place" (Meijer & Sijtsma, 2001, p. 130).

## Guiding Questions

In this dissertation I use the following questions to guide the study:

(1) What is differential person functioning?

(2) How do the methods for assessing differential person functioning differ across contexts?

(3) To what extent does differential person functioning contribute to our understanding of person fit across contexts?

Overall this dissertation will build upon previous research by delving deeper into the analysis of differential person functioning.

## Definitions

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

<u>Aberrant Response Pattern</u> – Person responses to a set of items that are not what would be expected given the model of analysis. The dominant research refers to this as an "aberrant" response. Within this study this response pattern will also be referred to as "unexpected" or "unusual".

<u>Differential Item Functioning (DIF) analysis</u> – A statistical procedure used to determine if items are an appropriate measure of an intended construct. The underlying question is as follows: Do items perform as intended for a given population? *DIF* occurs when individuals matched on the same latent variable have differing probabilities of endorsing an item.

Differential Person Functioning (DPF) analysis – A statistical procedure used to identify individuals who do not perform as expected on a set of items. The underlying question is as follows: Do persons perform as intended for a given set of items? *DPF* occurs when an individual's observed response pattern differs from the expected response pattern for individuals with the same location on the latent variable or construct. This is also referred to as an unusual or aberrant response pattern.

Erasure – Erasing one answer choice to choose another answer on a multiple-choice item.

Item Response Function (IRF) – The functional relationship between the probability of a correct response and the difficulty of an item. IRFs can be graphically depicted as a monotonically *increasing* ogive shaped curve whose slope changes as a function of the latent variable and difficulty of the item. This is also referred to as an Item Characteristic Curve (ICC).

Misfit – Refers to the inaccuracy of the approximate fit of a person response patter to a given model of analysis.

Operating Characteristic Function (OCF) – The functional relationship between the probability of a correct response and a logit scale. The person response functions and item response functions are defined based on how the x-axis is operationalized (Samejima, 1983).

Person Fit Indices – Measures of the degree of reasonableness of an individual's fit relative to a group of test items. The degree of misfit can be calculated with person fit statistics.

Person Response Function (PRF) – The functional relationship between the probability of a correct response and the achievement level of a given person. PRFs can be graphically depicted as monotonically *decreasing* ogive shaped curves whose slope changes as a function of the difference between person achievement and item difficulty. This is also referred to as a Person Characteristic Curve (PCC) and a Person Response Curve (PRC).

## Organization of Dissertation

Organization of the dissertation is as follows. Chapter One provided an introduction to the study including a statement of the problem, purpose of the study, and an outline of the questions guiding this research. Chapter Two addresses the first guiding question: What is Differential Person Functioning? The chapter covers a review of the literature that includes a discussion of person reliability, person response functions, person fit statistics, and response behaviors. Chapters Three and Four present the two case studies examined in the dissertation that illustrate the usefulness of DPF as a method for assessing validity of person scores. Both chapters provide a separate purpose, set of research questions, methods, results, and discussion. Finally Chapter Five draws connections among the studies while noting limitations. This chapter also provides readers with implications for research and practice.

# CHAPTER TWO: REVIEW OF LITERATURE

## DIFFERENTIAL PERSON FUNCTIONG

Through a review of the related literature, the evolution of differential person functioning is outlined in this chapter. The chapter is divided in themes, *Reliability of Person Measurements; Person Response Functions; Person Fit Indices;* and *Response Behaviors.* This chapter provides key research and theories for each theme. For ease of reference, a chronological list of important ideas from key researchers can be found in Table 1.

### Reliability of Person Measurements

Ideas of person invariance are not new and date back to early researchers, one in particular being Mosier (1940, 1941). The idea of person reliability or variability of a person on an assessment is of great importance. Person reliability is a theme evident across the works of many researchers (Keats, 1967; Lumsden, 1977, 1980; Mosier, 1940, 1941). One of the first mentions of person invariance in the literature occurred in the work of Mosier (1940, 1941) in the area of psychophysics. Mosier recognized that an individual's composite score may not be an accurate representation of an individual's location on a latent trait. Mosier posited that a person's score is dependent upon the person's variability with respect to the group of items used to obtain the score. Mosier (1940) came to recognize that the reliability of an assessment score is dependent upon the ambiguity of the assessment and "the variability of the individual" (p. 357).

In an early review of test theory, Keats (1967) expanded on Mosier's work in the area of person reliability and proposed solutions to eliminating the interference of individual characteristics on item responses. He focused on the necessity of ordering persons on ability and emphasized ensuring that test data satisfy this condition. For persons to be ordered, “each item would have to discriminate significantly between groups” (Keats, 1967, p. 218). Keats asserted that subjects should be grouped based on overall test scores rather than observed individually. However, later research comes to show that despite a group’s common test score index, response patterns can still be unique for each person and provide useful information for interpreting individual performance (Lumsden, 1977; Quaynor, Perkins, & Engelhard, 2009). Although Keats (1967) did not explicitly refer to differential person functioning, it is in fact what he addressed in his work. In other words, his concern was “whether or not subjects have been ordered with respect to more than one dimension” (p. 218).

Person variability was later explored in the areas of computerized adaptive testing (Vale & Weiss, 1975; Weiss, 1973) and on the Scholastic Aptitude Test (Levine & Rubin, 1979). Within the scope of computerized adaptive testing Weiss (1973) observed groups of individuals who correctly answered items of the same difficulty level yet differed on the items they answered correctly at different difficulty levels. Vale and Weiss (1975), also within the realm of computerized adaptive testing, investigated the premise that more consistent individuals would yield more stable ability estimates. In 1979, student response patterns to multiple-choice items on the Scholastic Aptitude Test were studied (Levine & Rubin, 1979). The populations considered by Levine and Rubin (1979) were individuals who obtained lower scores because of problems with English language fluency and individuals who obtained higher scores because of cheating. Levine and Rubin (1979) went on to develop numerical measures called appropriateness indices to identify these individuals. These measures and others will be considered in a subsequent section of the dissertation, *Person Fit Indices*.

Using psychological measurement and mental growth as the backdrop for his work, Lumsden (1977, 1980) provided a useful approach to address issues of person reliability. Lumsden (1977) presented what he referred to as an “attribute based model of test performance” (p. 477). In classical test theory, reliability is viewed as an aspect of group separation or variability. Lumsden proposed using this idea to suggest that person consistency or reliability should be examined when interpreting person scores. Not only did researchers like Lumsden (1977, 1980) recognize that person reliability was an important idea to study they also constructed ways for assessing the variability of individuals which include the use of graphical representations of person response patterns (Keats, 1967; Lumsden, 1977; Perkins & Engelhard, 2009; Trabin & Weiss, 1979; Vale & Weiss, 1975; Weiss, 1973).

## Person Response Functions

In 1973, Weiss proposed a graphical representation of the relationship between item difficulties and individual responses to items called a trace line. Weiss (1973) illustrated that given a set of items, as item difficulty increased, the individual’s percentage correct would decrease. Lumsden (1977) later elaborated on the idea of a trace line and introduced the use of a person characteristic curve (PCC). “The person characteristic curve is the plot for a single subject of the proportion of items passed at different difficulty levels. It is perfectly analogous to the item characteristic curve” (Lumsden, 1977, p. 478). Lumsden served as the first researcher to clearly define this term, although the idea was implicit in previous research (Keats, 1967; Vale & Weiss, 1975; Weiss, 1973). The underlying idea behind the PCC is that a person receives a correct response on an item when their location on a given latent variable is greater than the given location of an item. Person responses curves (PRCs), constructed using the same method as Lumsden (1977) for PCCs, were studied by Trabin and Weiss (1979). Trabin and Weiss (1979) compared expected and observed PRCs using vocabulary test responses of college students. In their study, Trabin and Weiss (1979) found that 90% of students fit the expected PRCs. The expected PRC served as a good predictor of observed PRCs for the population.

I will use the term person response function (PRF) henceforth to refer to the functional relationship between the probability of a correct response and item difficulty. As identified by previous research, PRFs can be graphically depicted as monotonically decreasing ogive shaped curves whose slope changes as a function of the difference between person achievement and item difficulty.

**Crossing Person Response Functions**

Through the appraisal of crossing PRFs, Lumsden (1977) identified issues associated with the use of total test score for grouping individuals when addressing issues of ordering persons. Lumsden examined situations of crossing PRFs in which two subjects received the same total test score but when they were examined in relation to their correct responses on items ordered by difficulty their PRFs differed and crossed. This crossing illustrated their differing response patterns. The crossing of PRFs results in the estimates of reliability being “biased by the difficulty of the items” (Lumsden, 1977, p. 481).

If person response functions of individuals with identical values of $\theta$ cross, this could be an indicator of differential person functioning (DPF). Perkins and Engelhard (2009) demonstrated the impact of crossing PRFs using the following example. Figure 1 illustrates the effects of crossing PRFs. Three PRFs are illustrated for two situations: Rasch PRFs, adhering to the requirements of invariant measurement, that *do not* cross (Panel A) and Birnbaum PRFs that *do* cross (Panel B). As shown in Panel C, non-crossing PRFs yield comparable person locations over subsets of items centered around easy items (-2 logits) to hard items (+2 logits). If PRFs do not cross, then Persons A, B, and C are ordered in the same way across item subsets (Lumsden, 1977). In other words, item-invariant measurement is achieved with the Rasch model.

Crossing PRFs based on the Birnbaum model (Panel D) yield person ordering that varies as a function of the difficulty of the item subsets. For example, Person A is the lowest achieving person with the lowest probability of success on the easy items, while Person C is the highest achieving person on the hard items. In this example, easy item subsets yield persons ordered as A < B < C, while hard item subsets yield persons ordered B < C < A. In other words, the ordering of persons is not invariant over item subsets with the Birnbaum model.

The ordering of persons below and above the intersection points vary when PRFs cross (Perkins & Engelhard, 2009). Crossing PRFs can lead to problems with the substantive interpretation of person performance (Lumsden, 1977, Perkins & Engelhard, 2009). Lumsden (1977) identified the importance of obtaining this diagnostic information. This data could bring about important information for teachers in particular when addressing instructional strategies for individual students.

PRFs provide a way to define, visualize, and analyze differential person functioning. The PRF is a function with the potential to detect unexpected response patterns (Sijtsma & Meijer, 2001). In order to evaluate the type of behavior attributable to the unexpected response pattern, person fit statistics can be used to gain a better understanding of the population and assessment under investigation. Engelhard (2009) emphasizes that the presence of DIF and DPF signify that the requirements of invariant measurement are not met by the item-person responses. He suggests analyses of residuals, standardized residuals, and mean square error statistics (Infit and Outfit) to examine person fit. The utility of this method is presented in his study of the assessment of students with disabilities (Engelhard, 2009). Additionally, he proposes the development of a mixed methods approach in psychometrics to address issues of individual performance.

## Person Fit Indices

Person response functions used in conjunction with person fit indices yield more information for scholars interested in studying student response behavior (Embretson & Reise, 2000). As noted, in some instances, the overall test score is not an appropriate measure of the construct for a given student (Levine & Rubin, 1979). Early researchers such as Levine and Rubin (1979) and Van der Flier (1982) investigated unusual response behaviors and developed statistical measures to quantify the reasonableness of response patterns. Levine and Rubin (1979) developed what they referred to as appropriateness indices. Appropriateness indices are measures of the fit of individual response patterns to psychometric models. Thus, this index is only a function of the examinee’s responses. A description of this index along with a cadre of other person fit indices discussed in this section can be found in Table 2. Appropriateness indices allow researchers to identify students who do not approach the test in the same way as other students with the same achievement level. Levine and Rubin (1979) saw these indices as useful measures for identifying “spuriously” high and “spuriously” low examinees. In 1983, Harnisch and Tatsuoka provided a review of fourteen appropriateness indices using 1977 NAEP data. In their work a variety of indices were investigated. It was found that while many of these indices were highly correlated quite a few were unrelated. The study was a comparative analysis of four groups of indices, extended caution indices, standardized extended caution indices, appropriateness indices, and item response model indices (Infit and Outfit). Harnisch and Tatsuoka (1983) found that appropriateness indices a generally related to the total score.

Van der Flier (1982) also examined student response patterns using what he refers to as deviant scores through the lens of cross-cultural psychology. He asserted that there are many problems when comparing the test scores of groups from differing cultural backgrounds. Van der Flier noted that if a test has some given meaning at a group level comparison, it doesn’t indicate that the same meaning holds at the individual level. Deviance scores represent how much an individual’s observed score pattern differs from an expected score pattern (Van der Flier, 1982). A high deviance score is likely an indication that the test score is not an accurate representation of the construct of interest. However, the meanings of deviance scores are specific to the test and population of interest. Generalizations of meanings to other contexts are problematic.

Person-fit research is defined as the use of “methods to identify respondents whose pattern of scores on the items from a test or questionnaire is unusual, given the expectation based on a particular item response theory model, or given the item score patterns produced by the majority of the respondents" (Sijtsma & Meijer, 2001, p. 191). One approach for identifying unusual response patterns is the use of person-fit indices to measure the degree of fit of an individual's responses to a group of test items. While PRFs provide a visual method for detecting unexpected behavior, person fit analysis can quantify this behavior.

In a study concerned with the systematic investigation of individual performance on assessments, Rudner (1983) evaluates nine proposed indices for assessing person fit. Two of the indices were based on the Rasch (1960) model, the unweighted total fit mean square (Outfit) and the weighted total fit mean square (Infit). Three indices were based on the Birnbaum (1968) model, the unweighted and weighted total fit mean square indices, calculated using a three parameter model, and a third approach based on the likelihood function $L(\theta_i)$. Rudner (1983) also evaluated two correlation coefficient approaches and two item sequencing approaches. Rudner suggested that the power of the application would be a possible criterion for selecting one proposed index over another. Other criteria included available item parameters and computation requirements. Additionally, the type of assessment might dictate the selection of the critical value which would provide a cut-off for determining statistical significance. Rudner suggests that for classifying misfit, large-scale assessments as opposed to classroom tests might be better suited to employ a conservative decision rule.

Masters (1988) approaches the idea of differential performance from the angle of the traditional discrimination parameter found explicitly in the 2PL and 3PL IRT models proposed by Birnbaum (1968). Discrimination is interpreted as the degree to which the item makes a distinction between individuals on the latent variable of interest. The discrimination parameter is also referred to as the slop parameter. Masters (1988) postulates that high discriminations, normally a desirable characteristic of an item within classical test theory, could be an indication of a measurement problem from the perspective of IRT. Highly discriminating items effectively discriminate between low achieving and high achieving students. He suggests that the Rasch model is unique in that it proposes items with high discriminations be eliminated from the test. As Keats (1967) suggests, items must discriminate between groups for persons to be ordered. The Rasch model motivates researchers to ask why an item has a high level of discrimination, this may indicate a problem in the assessment of an individual. Masters (1988) suggests that this differential item performance could be the result of a number of issues including opportunity to answer and testwiseness, a student's keenness at taking assessments. Further investigation of the discrimination parameter as it relates to persons could result in a better understanding of person sensitivity to items.

Another approach for assessing person fit was taken by Klauer and Rettig (1990) who proposed a standardized person test for assessing consistency with a latent trait or IRT model. They compared three Chi-Square based statistics, $\chi^2_{SC}$, $\chi^2_{W}$, $\chi^2_{LR}$, as options for evaluating the invariance hypothesis for tests of shorter lengths, i.e. less than 80 items. The statistics are computed using single response vectors that are standardized such that their conditional probabilities do not depend upon the absolute value of an individual's theta.

Reise (1990) investigates the proposition that a traditional item-fit index can be used to assess person fit (and vice versa). Through the comparison of a $\chi^2$ item-fit index with a likelihood-based person-fit index, Reise shows that while many item fit indices work well to identify misfitting items the same indices were not as successful with identifying severely misfitting individuals. Ultimately, Reise recommended a likelihood-based index to evaluate both examinee and item model-data fit.

There exists a plethora of comparative research on the quality of person-fit statistics (Karabatsos, 2003; Li & Olejnik, 1997; Reise 1990; Rudner 1983; Rudner, Bracey, & Skaggs, 1996; Sijtsma & Meijer, 2001; Smith, 1986). However, this research has not resulted in agreement as to which statistic is most useful given the characteristics of the test and person population of interest. For this reason, Karabatsos (2003) presents a comprehensive analysis comparing various parametric and non-parametric fit statistics under differing test conditions to ascertain a decision as to which person fit statistics is best suited for identifying unusual person response patterns and the virtues of each fit statistic. It is important to take a moment to note the difference between parametric and nonparametric item response theory models. Nonparametric item response theory models (NIRT) are based on rank ordering respondents. As defined by Sijtsma and Molenaar (2002), due to these order restrictions, any pair of $\theta$, such as $\theta_A$ and $\theta_B$, with $\theta_A < \theta_B$,

$$P_i(\theta_A) \leq P_i(\theta_B) \qquad [8]$$

While IRT models such as the Rasch model are parametric models "because they determine the relationship between $P_i(\theta)$ and $\theta$ by means of a parametric…function with scalar parameters" (Sijtsma & Molenaar , 2002, p. 13).

Ultimately, Karabatsos (2003) found five optimal person fit statistics out of the 36 that were investigated. The $H^T$ statistic, a non-parametric statistic, suggested by Sijtsma (1986) and Sijtsma and Meijer (1992), was found to be the best overall. It was determined that the $H^T$ statistic was not only the best with identifying students with unusual response patterns generally but also with detecting aberrant examinees on exams of varying lengths and examinees with a variety of different response behaviors.

Infit MNSQ and Outfit MNSQ previously discussed in Chapter One are the most popular fit indices investigated by researchers utilizing the Rasch measurement model. Many researchers have investigated the utility of these measures for assessing item and subsequently person fit (Harnisch & Tatsuoka, 1983; Karabatsos, 2003; Meijer & Sijtsma, 2001; Petridou & Williams, 2007; Rudner, 1983; Smith, 1986; Smith & Hedges, 1982). Smith and Hedges (1982) for example, correlated Infit and Outfit with a likelihood ratio fit statistic and found that they were highly correlated. They recommended the use of either fit statistics when assessing fit. Additionally, Smith and Hedges (1983) suggested the use of both Infit and Outfit to obtain the greatest amount of information regarding the distribution of the data.

The selection of critical values is often arbitrary when identifying "misfit". Researchers such as Bond and Fox (2007) and Wright, Linacre, Gustafson, and Martin-Lof (1994) suggest a cut-off of 1.3 which is often considered the conventional cut-off score for both Infit and Outfit. Other researchers have used simulation studies to acquire cut-off scores (Karabatsos, 2000; Petridou and Williams, 2007). We know that the distribution of Infit and Outfit statistics changes given the data set (Karabatsos, 2000).

This section provided a brief discussion of only a selection of person fit research that has been underway. Selecting which index is the best choice given the data and research questions of interest can be a daunting task. Meijer and Sijtsma (2001) suggests that when selecting which index is appropriate to use given the data and measurement model one should consider the fact that “detection rates are highly dependent on (1) type of misfitting response behavior, (2) θ value, and (3) test length.” (p. 130)

While person fit indices quantify unusual person response behavior, they do not provide explanations for such behavior. Various types of response behaviors have been identified by researchers as possible explanations for unexpected response patterns particularly for the assessment of the academic achievement of students (Hulin, Drasgow, & Parsons, 1983; Karabatsos, 2003; Meijer, 1996; Trabin & Weiss, 1979; Wright & Stone, 1979). In the following section I address research connected to these behaviors.

## Response Behaviors

Researchers have suggested a variety of response behaviors that could possibly lead to an inappropriate measurement of the construct for an individual (Hulin, Drasgow, & Parsons, 1983; Karabatsos, 2003; Meijer, 1996; Trabin & Weiss, 1979; Wright & Stone, 1979). These behaviors include but are not limited to sleepiness of the respondent, guessing, cheating by the respondent, inappropriate proctor assistance, lack of precision, and alignment error. Sleeping behavior can be the result of an examinee who needs time to get warmed up to the assessment resulting in incorrect answers of initial easy items and a higher percentage of correct answers on more difficult items. Sleeping behavior for individuals at higher locations on the attribute continuum (variable map) can possibly be identified when unexpected errors are found at the start of the assessment. It has been found that these individuals exhibit high Outfit MNSQ values (Linacre, 2009.)

An individual exhibiting guessing behavior would likely be at a lower achievement level (Linacre, 2009). Such a student would answer items of low and medium ability correctly while receiving a higher proportion of more difficult items incorrect. Typically individuals with lower locations on the attribute continuum who guess will have high values of Outfit MNSQ (Linacre, 2009). Cheating behavior is greatly associated with a low achieving student (Linacre, 2009) who would be expected to receive items of low difficulty correct and higher difficulty wrong but in fact receives a greater than expected number of higher difficulty items correct (Meijer, 1996). A student who works very methodically, slowly, and precisely could generate a score pattern resembling the Guttman (1950) model in that if the items are ordered on difficulty when a student responds correctly to a particular item they will also respond correctly to all items of lesser difficulty. However, person responses are typically probabilistic in nature and as such a Guttman response pattern would be unusual.

Hulin, Drasgow, and Parsons (1983) suggested alignment errors as a possible source of unexpected behavior. In this case, a multiple-choice exam would be administered with a test form in addition to a separate answer sheet. A student with a high achievement level would have an unexpectedly high proportion of incorrect responses on both low and high difficulty items. This could be the result of alignment errors when recording answers to the answer sheet.

Another possible response behavior is one where the student incorrectly answers many easy items but answers more difficult items correctly. Upon first glance this might appear as a case of cheating. However, if the items of lower difficulty all represent a particular sub content area it could be the case that the student has yet to master a skill set that was assumed to have been learned. The patterns of response discussed here are merely suggestions of suspicious behavior, and further analyses would be necessary (quantitative and qualitative) to accurately assess the response behavior of a given individual. In some cases accommodations are made for these response behaviors. For example, Cronbach (1946) suggested the use of specialized scoring keys to weigh answer choices differently for different response behaviors or in some cases simply invalidating the student score. An approach to dealing with guessing on assessments is to make an adjustment to the measurement of the student score. Yet, as Smith (1986) pointed out, this correction "has been applied blindly" with little to no concern taken to the pattern of guessing encountered whether right answers to hard items or right answers to easy items (p. 361).

Trabin and Wiess (1983) provide graphical descriptions of PRFs for individuals based on their response behavior. For example, when an individual has correctly answered questions above their ability level, graphically displayed by a dip in the curve, it is assumed that this individual likely guessed on this question as they likely have not acquired the appropriate knowledge level to choose a correct response otherwise. Another behavior which can be determined graphically would be carelessness. If an individual has a large percentage of incorrect responses to items located below their ability level it is safe to infer that the individual is displaying some level of carelessness. This example provided by Trabin and Weiss (1983) is not meant to generalize across all individuals or all assessments. It does however serve as an illustration of the possible utility of PRFs in understanding and detecting unique response behaviors.

Unusual responses can also be an indication of a student's opportunity to learn including access to necessary supplies and the presentation of instructional materials. As discussed, there exist response behaviors that can impact the person variability. Person variability can be examined through the use of graphical representations (Keats, 1967;

Lumsden, 1977; Perkins & Engelhard, 2009; Trabin & Weiss, 1979; Vale & Weiss, 1975; Weiss, 1973).

The historical evolution of differential person functioning is tracked in this chapter through an extensive review of the literature, showing that person reliability has been a concern of researchers over the last century. The chapter also highlights the utility of person response functions in identifying and understanding unexpected response patterns. However, it is important to understand that PRFs should be utilized in conjunction with person fit indices in assessing student response behaviors (Embretson & Reise, 2000). The evolution of person fit indices in the last 30 years was also explored from Van der Flier's (1982) deviant scores to Sijtsma and Meijier's (1992) $H^T$ statistic. We see from the review of the literature that there are many proposed person fit indices.

In this dissertation I have chosen to incorporate the traditional Rasch based fit statistics of Infit and Outfit in exploring invariant measurement through the lens of DPF. The following Chapters (three and four) contain the two case studies used to provide illustrations of investigations of DPF.

# CHAPTER THREE: CASE STUDY ONE

# USING PERSON FIT TO EXAMINE ERASURE DATA

This chapter presents the first of two case studies within this dissertation. This case study approaches the detection of differential person functioning through an exploration of the relationship between wrong-to-right erasures, person fit indices, and school-level mathematics and reading achievement using a pre/post erasure design. The chapter details the research questions that were assessed, data used in the study as well as the methods and results followed by a brief discussion which will be elaborated upon in Chapter Five. A summary of this information is found in Table 3.

## Introduction

Student erasure practices (erasing one answer choice to choose another answer on a multiple-choice item), which have caused challenges in urban school districts, are one threat to the validity of assessments. Erasures can happen for a variety of reasons, such as the rethinking of an item by a student, misalignment of the answer sheet used by a student, or improper assistance in modifying item responses from an outside source (Mead, Anderson, & Korts, 2010). As pointed out by Qualls (2001), “it is possible through an examination of erasure behavior to determine what is typical behavior and to begin to use it to flag deviant patterns” (p. 10). Unexpected response patterns may no longer be an accurate and fair representation of student knowledge, and therefore should be identified for further investigation. Findings from erasure analyses have emerged as an indicator of potentially unethical behavior by teachers and administrators. Current studies range from research on the answer changing behaviors of students (van der Linden &

Jeon, 2012) to studies that explicitly focus on erasure analyses as a key component for detecting teacher cheating (Amerin-Beardsley, Berliner, & Rideau, 2010).

Amerin-Beardsley, Berliner and Rideau (2010) have gone as far as to relate cheating by school personnel to severe criminal offenses ultimately classifying cheating into three categories of offenses $1^{st}$, $2^{nd}$ and $3^{rd}$ degree. They consider the erasing of student answer responses as a $1^{st}$ degree cheating offense. The charges made against teachers and school administrators are serious, and unfortunately strong evidence has surfaced that this behavior is not occurring in a vacuum. This study provides useful information for school districts and administrators concerned with assessing unusual response patterns, in particular irregular erasures.

## Purpose

The occurrence of erasures is not an issue. The problem lies in distinguishing between regular and irregular erasures. In recent years, erasure practices that indicate educator cheating have dominated media conversations surrounding education and high stakes testing. Given the high profile of this topic, the importance of adequate and robust techniques for examining erasure behavior is very important. This study builds on the foundation laid by researchers in the last 30 years related to analyses of erasure behaviors by including person fit research in the investigation of erasures. Irregular erasures can impact the validity of person scores. Combining person fit analysis with the examination of erasure analysis allows for a quantitative appraisal of differential person functioning. The purpose of this case study is to illustrate the usefulness of differential person functioning as a method for assessing the validity of person scores through an exploration of the relationship between erasure behavior, person fit indices, and school-level achievement.

## Research Questions

This case study utilizes mathematics and reading achievement data from a statewide standards-based assessment to explore the following research questions:

(1) Is there a relationship between wrong-to-right erasures and mathematics and reading achievement at the school-level?

(2) Does the relationship between wrong-to-right erasures and mathematics and reading achievement vary based on school context?

(3) Is person fit a useful index for detecting irregular erasure behavior at the school level?

## Methods

### Data

Data was obtained from a statewide standards-based assessment given to students in a northeastern state in the United States of America. Response patterns for students on the 2010 administration of the mathematics and reading sections of the assessment are analyzed.

Given that the data are from a secondary source, there were some constraints in data manipulation. To effectively examine the data, data management was approached in three steps. In step one data were obtained in the form of two files, the erasure file and the item file. The erasure file contained rows representing each erasure made by a student. Therefore students may have multiple or no entries in the file. The columns represented the types of erasures a student could have performed: wrong-to-right (WR), right-to-wrong (RW), and wrong-to-wrong (WW). The item file contained a row for each student in the data set. While the columns indicate item responses and student demographic information along with school and district affiliation.

In step two decisions were made to narrow the focus to one grade level and one form allowing for a more manageable data set. This also allowed for the erasure and item files to be combined without complication. Grade 3 and form H were chosen (n=4,268) for analysis in the dissertation. The grade 3 assessment contained 28 forms. Form H was chosen using a random number generator in Microsoft Excel which excluded special forms (Braille, Large Font, etc.). Approximately half the sample is female (48.3%) (Table 4). White, Asian, Hispanic, and African American ethnicity groups are represented at the following rates 55.40%, 19.12%, 13.61%, and 11.05% respectively. Data is provided for 48 schools within 22 districts.

Because the data provided student response patterns and their corresponding erasure behavior (None, WR, RW, and WW), it is possible to infer the student response patterns before an erasure occurred. These inferred response patterns will be called pre-erasure strings. For an illustration refer to Figure 2. Here you see the erasure behavior for one student on a set of ten items. This student erased only their responses to items 4 and 7 with these erasures being from a wrong choice to a correct choice, wrong-to-right (WR). This is reflected in the constructed pre-erasure string which illustrates that had this student not erased on these items the answers would have been incorrect. In step three of the data management processes pre-erasure strings were constructed for every student on the assessment, with only WR erasure behavior considered in the creation of the pre-erasure strings. SPSS 19 software was used to create the pre-erasure strings and combine the resulting item and erasure files.

This study focuses on the 35 and 18 multiple choice items within the mathematics and reading sections of the assessment respectively.

*Mathematics*

The mathematics section contains a total of 44 items, 35 multiple choice, 6 short constructed response, and 3 extended constructed response items. This included four content areas, number and numerical operations; geometry and measurement; patterns and algebra; and data analysis, probability, and discrete mathematics. The short constructed response items were holistically scored on a scale of 0 to 1 and the extended constructed response items were holistically scored on a scale of 0 to 3. Students were able to earn a maximum score of 50 on the grade 3 mathematics section. Analysis will only occur on the 35 multiple choice items.

A conceptual model for this study within the content area of mathematics is presented in the upper panel of Figure 3. This model depicts the construct, mathematics achievement which is made observable by the 44 items. The dashed line represents construct-irrelevant variance in the form of possible student and school-level factors.

*Reading*

Within the language arts and literacy section of the assessment there are two clusters, reading and writing. Only the reading cluster will be examined in the dissertation. The reading cluster consists of 21 items, 18 multiple choice and 3 constructed response items. Just as for mathematics, analysis will only occur on the multiple choice items. Items are grouped based on two skill areas, working with/interpreting text and analyzing/critiquing text. Additionally, the reading passages include literature (narrative) readings and everyday (informational) readings. The constructed response items were holistically scored on a scale of 0 to 4.

The conceptual model for the reading content area is very similar to the conceptual model presented for the mathematics content area. Both can be viewed in Figure 3. In the bottom panel, you can see that the latent construct, reading achievement, is made observable by the 21 items.

Data used in this dissertation were obtained following Institutional Research Board (IRB) guidelines for my institution and the State Department of Education where the data were collected (Appendix A).

**Study Design**

As discussed in Chapter One, Rasch (1960/1980) measurement theory allows for the development of assessments that adhere to the requirements for invariant measurement as set forth by Engelhard (2013). Recall, Equation 1 which illustrates the Rasch model for dichotomous responses. This traditional model has two facets, persons and items. This case study is concerned with multiple facets and thus utilizes the Many Facets Rasch model (MFRM) which allows for multiple facets to be examined (Equation 8). I used the Facets computer program (Linacre, 2010) to analyze response data with the MFRM.

$$\phi_{nimj} = \frac{P_{nimj}}{P_{nimj-1} + P_{nimj}} = \frac{\exp(\theta_n - \delta_i - \Delta_m - \mu_j)}{1 + \exp(\theta_n - \delta_i - \Delta_m - \mu_j)} \qquad [8]$$

where,

$\theta_n$ = location of a student on the latent variable (mathematics/reading achievement)

$\delta_i$ = difficulty or location of the item

$\Delta_m$ = Pre/Post erasure identifier

$\mu_j$ = school

This model allows for the analyses of several facets, in particular this study analyzes, students, items, pre/post identifier, and schools. The MFRM is used to calculate person fit statistics. This case study is concerned with Infit MNSQ and Outfit MNSQ as well as standardized Infit and standardized Outfit. These statistics were discussed in more depth in Chapter One. They were chosen for analysis in this study for three reasons, 1. I used a Rasch based model to examine the data and these are traditional Rasch fit statistics, 2. The mathematics and reading assessments investigated in this study were constructed with Rasch based models, and lastly 3. Infit and Outfit have been found to be promising statistics for obtaining person fit information. Through the use of person fit statistics from MFRM analyses, variable maps, and erasure indices the research questions are examined.

In addition to using a MFRM to assess the data I also made use of a pre/post erasure design. This study specifically examines wrong-to-right erasure behavior. Pre-erasure strings were constructed for each student. The pre-erasure strings take into account the expected response string if no wrong-to-right erasures occurred. This method of data analysis was adapted from Mead, Anderson, and Korts' (2010) analyses of erasures and Rasch residuals. Given this design choice, small increases in achievement are expected when comparing pre and post erasure response strings, large increases in achievement may suggest a large proportion of irregular erasures. Another design choice made for this study was to focus on the school as the level of analysis.

# Results

In this section, results are presented for each content area. Within each subsection baseline information of school-level erasure behavior is provided. Then the relationships between wrong-to-right erasure behavior and mathematics and reading achievement are examined. Lastly, a comparison was conducted based on content area.

## Mathematics

Of the 6,691 erasures in mathematics, 66.19% were wrong-to-right (WR) erasures. Table 5 presents percentages and means of school-level erasure behavior providing. The percentage of WR erasures within schools ranges from 41.59% to 89.90% (School 903) and the mean WR erasures per student ranges from 0.59 to 3.48 (School 908).

Using the Facets computer program (Linacre, 2010) the MFRM was applied to the data (Table 7). Facets summary statistics indicate values of 1.0 for Infit MNSQ and Outfit MNSQ which suggest that there was minimal misrepresentation in the measurement system used to establish the assessment. The summary statistics were analyzed in conjunction with the variable maps (Figures 4 and 5). Variable maps allow for the visualization of the latent variable, in this case mathematics achievement, on a continuum where locations at the top of the continuum signify a higher level of mathematics achievement and locations at the bottom of the continuum signify a lower level of mathematics achievement. As you can see the variable map allows for a display of students in terms of their “ability” and items in terms of their difficulty on the same scale. The facets summary statistics and variable maps indicate that the reliability of separation for persons is quite good at .90. There is good separation of the items in terms of defining the variable. There is significant variation in schools on their levels of achievement. On average there was a small increase in student achievement from the pre to post erasure item responses.

In further identifying baseline information for the sample on erasure behavior by school, the mean WR erasures by total erasures was examined in Figure 8. This information allows for a visual representation of the spread of mean WR erasures across schools. Schools 903 and 908 have slightly higher mean WR erasures than the other schools in the sample.

School level person fit statistics were obtained from the Rasch analysis based on the pre and post erasure response strings (Table 9). Based on this information there were no schools indicated with unreasonable Infit MNSQ or Outfit MNSQ values, values outside of the range of 0.8-1.20. Figure 9 provides baseline information for the relationship between mean pre and mean post school achievement indicating a strong positive direct correlation. The interaction between school-level mathematics achievement and the pre/post indicator reveals that there is a larger variation in school achievement for schools 903 and 908 than the other schools in the sample (Figure 10). Outfit Z and a *Z* statistic were examined in their relationship with wrong-to-right erasures. Outfit Z, standardized Outfit tests the hypothesis of whether the date fit the model perfectly. Based on the observations of Outfit Z in Table 9, the data for the schools have reasonable predictability, that is the values are within the range of -1.9 to 1.9. The Z statistic is the aggregate standardized residual for each school. Neither Outfit Z nor Z demonstrated much variation at the school-level. However, the relationship between these person fit statistics and wrong-to-right erasures for schools 903 and 908 were unlike other schools in the data (Figures 11 and 12). Lastly, the relationship between *Z* calculated using pre erasure data and *Z* calculated using post erasure data showed a direct positive relationship (Figure 13).

**Reading**

Of the 2,772 erasures in the reading content area, 56.39% were WR erasures. The percentage of WR erasures within schools ranges from 25.00% to 100.00% (School 922) and the mean WR per student ranges from 0.16 to 1.13 (School 908) (Table 6).

Facets summary statistics indicate values of 1.0 for Infit MNSQ and Outfit MNSQ as in the Mathematics section, suggesting that there was minimal misrepresentation in the measurement system used to establish the assessment. Facets summary statistics indicated that the reliability of separation for persons is good at .83 (Table 8). There is good separation of the items in terms of defining the variable. There is significant variation in schools on their levels of achievement with a small average increase in student achievement from pre to post erasure item responses. The variables maps suggest that the items might have been on average easier for the students (Figures 6 and 7).

In examining the mean WR erasures by total erasures at the school level, school 908 had slightly higher mean WR erasures than the other schools (Figure 14).

Figure 15 provides baseline information for the relationship between mean pre and mean post school achievement. This positive direct correlation relationship is as expected. School 908 appears to stand out from the pact. The interaction between school level reading achievement and the pre/post indicator does not reveal that there is a larger variation in school achievement for any one school (Figure 16). Neither Outfit Z nor the

*Z* statistic demonstrated much variation at the school-level. However, the relationship between these person fit statistics and wrong-to-right erasures for school 908 was somewhat unlike other schools in the data (Figures 17 and 18). The relationship between *Z* calculated using pre erasure data and *Z* calculated using post erasure data showed a direct positive relationship with school 908 showing some variation (Figure 19).

### Comparison

There are less total erasures in the reading content area than mathematics which is due to the difference in the number of items in each section. The reliability of separation is lower in the reading content area. This once again could be a result of the lower number of items. Both content areas show a significant variation in schools on their levels of achievement. In both content areas, on average there was a small increase in student achievement from pre to post erasure item responses. The variable map for reading shows students higher than items unlike the mathematics content areas, suggesting reading items are easier for this population of students. Outfit Z is less in reading, and this is likely because items are nested within passages.

Baseline information of schools in both content areas indicates some schools that warrant further investigation. Schools 903 and 908 stand out in the mathematics content area as schools which should be studied further, while schools 922 and 908 stand out in the reading content area. In particular, the analyses suggest that School 908 should be studied further.

## Discussion

Studies on erasure behavior are particularly important given that “the many behaviors that constitute cheating combine to diminish our ability to accurately gauge student achievement" (Cizek, 2003, p. 31). This failure to accurately gauge student achievement distorts our ability to interpret the meaning of the test scores, which as we know can affect the validity of the assessment system (Cronbach, 1971).

This case study is an explorative study that presents a method for detecting erasures and thereby providing baseline information enabling irregular erasures to be identified. When interpreting the results a limitation to the analysis should be kept in mind. In creating the pre erasure strings only the wrong-to-right erasures are taken into account. This does not account for all the erasures a student made and subsequently eliminates a portion of the erasures made by a student. This study design was chosen because the assumption was made that wrong-to-right erasures are an appropriate and sufficient indication of irregular behavior in this context.

Ultimately in addressing the research questions posed in this case study, I found that there is a relationship between wrong-to-right erasures and mathematics and reading achievement at the school level. On average there was an expected small increase in student achievement as the quantity of wrong-to-right erasures increased. The estimates of school achievement for pre and post erasure seem to be promising as a way to identify schools with irregular erasures (Research Question #1). It does appear that the relationship between wrong-to-right erasures and mathematics and reading achievement vary based on school context. This is evident based on the graphical depictions and correlations of erasure statistics and person fit statistics. However, some variation appears to be due to the structural differences in the test sections (Research Question #2). Finally, person fit (Infit MNSQ, Outfit MNSQ, Outfit Z and Z statistics) did not seem sensitive to the pre and post erasure changes at the school level (Research Question #3).

This study displays the usefulness of differential person functioning as a method to assess invariant measurement by assessing the validity of person scores through an investigation of erasure analysis and person fit analyses. In summary, this study suggests the pre/post erasure design has the ability to provide useful information for stakeholders concerned with the accuracy of person scores. It is also clear that more research should be conducted utilizing this study design to garner more information of its power in detecting and understanding erasures at a student and school level.

Lastly, it is important to remember that statistical analyses of erasure patterns cannot provide conclusive proof for any decision and inferences about inappropriate behaviors, such as adults changing student responses (Qualls, 2001, p. 10). Erasure analyses should never be used blindly or relied upon as the sole source of evidence for cheating. Instead these types of analyses can provide valuable information in erasure investigations.

# CHAPTER FOUR: CASE STUDY TWO

## USING A MULTILEVEL MODEL TO EXAMINE PERSON FIT

This chapter presents the second of two case studies within the dissertation. This case study examines the influence of a set of covariates (proficiency level, economic status, gender, and erasure behavior) on person fit within a multilevel framework. Person fit indices are analyzed with hierarchical generalized linear models for dichotomous data as a method to explore the usefulness of differential person functioning as a method for assessing the validity of person scores.

## Introduction

Decisions regarding students are made every day based on test performance. It is paramount that these data, in particular the student responses to test items, are accurate and support valid interpretations and uses of the test scores. Many researchers have identified factors that can hinder the accuracy of such student data (Hulin, Drasgow, & Parsons, 1983; Karabatsos, 2003; Meijer, 1996; Petridou &Williams, 2007; Trabin & Weiss, 1979; Wright & Stone, 1979). One impediment is that of unusual responses to items due to issues such as guessing, cheating, and carelessness. These unusual response patterns are defined as responses to a set of questions that are not what would be expected given a model of analysis. Individuals who display unusual response patterns have resulting test scores that are at risk for being measured inaccurately. Identification of this phenomenon is paramount for assessing the validity of test scores. As discussed in previous chapters person fit analysis represents one method in which these occurrences can be detected (Karabatsos, 2003). Person fit research typically only identifies that unusual responses exists and not the reasons why they exists (Meijer, 1996), leading researchers to speculate as to the cause(s) of unusual responses which can include demographic, behavioral, and organizational characteristics.

**Demographic and Behavioral Influences on Unusual Response Patterns**

Continuing on the path set forth by previous researchers, this study considers the relationship between student demographic variables and student response behavior. Petridou and Williams (2007) found that students who spoke more than one language at home, were less anxious, less motivated and more able in mathematics were statistically significantly more likely to have unusual  or aberrant response patterns. Many studies have also investigated the relationship of gender and ethnicity with unusual response patterns finding no significant relationship (Miller, 1986; Rudner, Bracey & Skaggs,1996; Petridou & Williams, 2007). This study considers the influence of gender on the likelihood of unusual response patterns. Potential significant findings would suggest that there are factors affecting one population of students that are impacting the validity of the assessment for them. Similarly, a student’s economic status could have a significant relationship with the likelihood of unusual response patterns. Implications from such a finding would lead one to consider the unique factors faced by an impacted population.

Achievement has been identified by researchers as one of the primary behavioral variables associated with unusual response patterns. Specifically, mathematics proficiency, and its relationship with person fit has been explored by researchers with mixed results (Chatman, 1985; Dodden & Darabi, 2009; Rudner, Bracey & Skaggs,1996; Petridou & Williams, 2007). Understanding the relationship between proficiency and the likelihood of unusual response patterns can aid in understanding student groups. For instance higher achieving students could be found to have a greater likelihood of unusual responses due to carelessness. While lower achieving students could have greater confusion on questions and perhaps more guessing on items resulting in a greater likelihood of unusual responding.

Response behaviors, behaviors that impact how students respond to test questions, which include student guessing, cheating, sleeping, etc., can impact student response patterns and can possibly lead to inaccurate measurements for a student on an assessment (Hulin, Drasgow, & Parsons, 1983; Karabatsos, 2003; Meijer, 1996; Trabin & Weiss, 1979; Wright & Stone, 1979). One behavior in particular, student erasure practices, defined as a student erasing one answer choice for another could have a significant relationship with student misfit. Erasures can happen for a variety of reasons not limited to student cheating, instructor or moderator interference, and misalignment of the answer sheet (Mead, Anderson, and Korts, 2010). Erasure behavior comes in three categories, wrong-to-right, right-to-wrong, and wrong-to-wrong. Each of these types of erasures in addition to the cumulative impact of these erasure types provides unique information about a student's performance. In the present study total erasures, all the erasure types a student might perform, are considered as a covariate, thus quantifying student erasure behavior in one variable. It is hypothesized that students increased erasure behavior will be associated with an increase in the likelihood of unusual responses. While each of these covariates; gender, economic status, proficiency level, and erasure behavior, alone can have significant influences on student response patterns the confounded affect of these factors must also be considered.

### Organizational Influences on Aberrance

In an effort to move beyond individual level influences on student response behavior, researchers have begun to investigate the impact of institutional and societal influences within multilevel frameworks. The last decade has seen an emergence of the use of multilevel models to study unusual response patterns (Conijn, Emons, van Assen, & Sijtsma, 2011; Petridou & Williams, 2007; Reise, 2000). Given the organizational structure of the educational system: pupils within classrooms – classrooms within schools – schools within districts, a multilevel approach to analyzing educational data is logical given that this type of model takes into account the hierarchical structure that exists. For example, district level policies can impact school level practices and school level practices can impact classroom instructional strategies. Each of these levels or even one of these levels can have a significant relationship with student misfit. This study extends the research by Petridou and Williams (2007) which suggests the classroom make a significant contribution to student aberrance by investigating between-school variation in aberrant responding. Specifically, it is hypothesized in this case study that the school a student attends impacts the likelihood of the occurrence of aberrant responses for that student.

## Purpose

The purpose of this study is to examine the influences of proficiency level, economic status, gender, and erasure behavior on person fit within the context of a high stakes assessment of mathematics and reading. If such influences exist they threaten the assessments adherence to the requirements for invariant measurement. As such this case study illustrates the usefulness of differential person functioning for assessing the validity of person scores.

## Research Questions

This study extends current literature on the person fit of student response data by examining student and school factors that may be associated with the likelihood of the occurrence of aberrant response patterns The four covariates considered are mathematics/reading proficiency levels, economic status, gender, and student erasure behavior. In choosing which covariates to include in the present study, variables were selected based on the current literature in person fit analysis and assessment validity taking into account the availability of variables in the data set. Specifically, the following research questions are explored:

(1) Is there significant between-school variation in the likelihood of the occurrence of an aberrant response pattern?

(2) Do select student- and school-level factors predict aberrant responding?

## Methods

### Data

Response patterns of grade 3 students (N=4,248) on the mathematics and reading sections of a 2010 statewide high-stakes assessment are utilized in this study (Table 4). This study focuses on the 35 and 18 multiple choice items within the mathematics and reading sections of the assessment respectively. Just as in the first case study, a conceptual model can be found in Figure 3 illustrating the constructs, mathematics achievement and reading achievement which are made observable by the items in each assessment. The dashed line represents construct-irrelevant variance in the form of possible student and school-level factors. Please refer to the "Methods" section of Chapter Three for a more detailed description of the data.

**Study Design**

Analyses are conducted in two steps. The first step employs a Rasch model to compute the fit statistics which are used as outcome variables. The second step considers a multilevel model to evaluate between-school variation in the likelihood of the occurrence of aberrant response patterns and the influence of the covariates on the likelihood of the occurrence of unusual response patterns. The analyses performed to address each research question in this case study are summarized in Table 11.

*Step One (Rasch Model)*

The Rasch model (1960/1980) for dichotomous variables using two facets was used to obtain the item and person parameters for the calculation of person fit statistics (Recall Equation 1).

Students with unexpected response patterns are identified using the Rasch mean square error fit statistics (MSE): Outfit and Infit. The analyses reported are based on a critical value of 1.20 for the Outfit and Infit statistics discussed in Chapter One. Therefore students with an Outfit or Infit value greater than or equal to 1.20 have responses that are classified as unusual/aberrant (Table 12). These fit statistics are then dichotomized to be used as dependent variables in the multilevel models within step two. Recall the discussion of Rasch fit statistics in Chapter Two for greater description of the calculation and rationale for the selection of Outfit and Infit. However, a few facts are important to note.

- Outfit is a measure more sensitive to outliers, responses where item difficulty is further away from an individual on a continuum where they share the same scale of measurement.
- Infit is a measure more sensitive to inliers, responses to items that are in line with an individual, i.e. on target.
- The expected value for Outfit and Infit is 1.0, with a reasonable range of 0.8 to 1.2.
- Values of Outfit or Infit that are less than 1.0 often indicate that the data is too predictable, resembling a Guttman pattern.
- Values of Outfit or Infit that are greater than 1.0 often indicate unpredictability of the model and that data is under fitting the model.

*Step Two (Hierarchical Generalized Linear Model)*

In step two of the analysis a hierarchical generalized linear model (HGLM) is applied to the data where the Infit and Outfit statistics are dichotomized and modeled as outcome variables, where 1 refers to aberrant and 0 refers to non aberrant. Three models are employed, an unconditional model (Model A), a model which includes the demographic variables (Model B) and one that includes the demographic and behavioral variables (Model C). Each of these models is conducted in four situations, when Outfit is the outcome variable and when Infit is the outcome variable within the mathematics and reading content areas. The models are as follows:

<u>Model A: An unconditional model</u>

Model A has no student- or school-level predictors and is referred to as the unconditional model.

Let $Y_{ij}$ take on a value of unity if the test responses of students $i$ in school $j$ display an aberrant pattern, with $Y_{ij} = 0$ if not; and $\mu_{ij}$ denote the probability $Y_{ij} = 1$. This probability varies randomly over schools. However, conditioning on this probability, we have

$$Y_{ij} \mid \mu_{ij} \sim B(m_{ij}, \mu_{ij}) \quad [9]$$

$$E(Y_{ij} \mid \mu_{ij}) = \mu_{ij} \qquad Var(Y_{ij} \mid \mu_{ij}) = \mu_{ij}(1-\mu_{ij}) \quad [10]$$

Here in level-1 of the HGLM $\eta_{ij}$ is the log-odds of the probability that the test responses of students $i$ in school $j$ display an aberrant pattern (Equation 11). This level-1 model accounts for the variation among students within schools.

$$\eta_{ij} = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \beta_{0j} \quad [11]$$

The level-2 model is

$$\beta_{0j} = \gamma_{00} + v_{00} \qquad v_{00} \sim N(0, \omega) \quad [12]$$

where $\beta_{0j}$ is the average log-odds of aberrant responding across schools and $\omega$ is the variance between schools in school-average log-odds of aberrant responding.

<u>Models B and C: Models with Predictors</u>

Student- and school-level factors are entered into Equation 8 to predict the likelihood of the occurrence of aberrant response patterns. Four explanatory variables were selected for inquiry in this study; two are categorized as demographic variables (gender and economic status) and two as behavioral variables (mathematics/reading proficiency and erasure behavior). Definitions of these explanatory variables can be found in Table 13.

Model B

In Model B the demographic variables are added to the unconditional model resulting in a level-1 model of

$$\eta_{ij} = \beta_{0j} + \beta_{1j}(Gender)_{ij} + \beta_{2j}(EconStat) \qquad [13]$$

where the gender and economic status variables are grand-mean centered. Thus $\beta_{0j}$ represents the average log odds of the occurrence of an aberrant response pattern. $\beta_{1j}$ represents the gender difference in the log odds of an aberrant response pattern, controlling for economic status. $\beta_{2j}$ captures the relationship between economic status and the outcome, holding constant gender. The level-2 model can be represented as

$$\beta_{0j} = \gamma_{00} + v_{00} \qquad [14]$$

$$\beta_{1j} = \gamma_{1j}$$

$$\beta_{2j} = \gamma_{2j}$$

Model C

The level-1 model for Model C includes all covariates, demographic and behavioral contrasting the advanced and partially proficient mathematics/reading achievement levels. Equation 15 represents this model for the mathematics content area.

$$\eta_{ij} = \beta_{0j} + \beta_{1j}(Gender)_{ij} + \beta_{2j}(EconStat) + \beta_{3j}(Erasures) + \beta_{4j}(MathAdv) + \beta_{5j}(MathP\Pr of) \qquad [15]$$

$\beta_{3j}$ =the relationship between log-odds of the occurrence of an aberrant response pattern and the total number of erasures, all else being equal

$\beta_{4j}$ =the difference in the log-odds of the occurrence of an aberrant response pattern between the advanced proficient mathematics group and the proficient mathematics group, all else being equal.

$\beta_{5j}$ =the difference in the log-odds of the occurrence of an aberrant response pattern between the partially proficient mathematics group and the proficient mathematics group, all else being equal

The level-2 model is

$$\beta_{0j} = \gamma_{00} + v_{00} \qquad [16]$$

$$\beta_{1j} = \gamma_{1j}$$

$$\beta_{2j} = \gamma_{2j}$$

$$\beta_{3j} = \gamma_{3j}$$

$$\beta_{4j} = \gamma_{4j}$$

$$\beta_{5j} = \gamma_{5j}$$

## Results

Using the Facets computer program (Linacre, 2010), the Rasch model was applied to the data to obtain the person fit statistics, Outfit MNSQ and Infit MNSQ. The Facets summary statistics (Table 14) indicate that the mathematics and reading assessments are likely well-constructed with little to no issues with the measurement system employed. The values for Outfit and Infit are within reasonable range with the Outfit MNSQ values for students and items in the Reading section slightly lower than the expected value of 1.0 at 0.98. This is a possible indication that the data runs the risk of over fitting the model.

Table 15 provides means and standard deviations for the covariates at each level (student and school). The means for student and school do not vary much however the standard deviations for students are greater than schools. The percentages of misfitting students by gender, economic status, and proficiency levels are provided in Table 16.

Henceforth, results are presented for two outcome variables, where $\text{Outfit}_A$ refers to Model A conducted with Outfit as the dependent variable, with similar nomenclature for all subsequent models. Results are presented for each content area, mathematics and reading. Multilevel model analyses were performed using HLM 7.0 software (Raudenbush, Bryk, Cheong, Congdon, & Du Toit, 2011). All results can be found in Tables 17 through 20.

**Mathematics**

Outfit

The results of the unconditional model ($\text{Outfit}_A$) suggest there is significant variation ($\tau$=0.051, $\chi^2$=69.038, p<.05) in the log-odds of an aberrant test response pattern on the Outfit variable at the school level (Table 17). The results of $\text{Outfit}_B$ indicate that there is a significant gender effect ($\beta$=-0.209, se=0.101, p=0.039). Female students are more frequently associated with a decrease in the occurrence in the log-odds of an aberrant pattern, holding constant economic status. In model $\text{Outfit}_C$, there was a statistically significant difference in the log-odds of the occurrence of an aberrant response pattern between the advanced proficient students and the proficient students ($\beta$=1.397, se=0.131, p<0.001) all else being equal. There was a statistically significant difference in the log-odds of the occurrence of an aberrant response pattern between the partially proficient students and the proficient students ($\beta$=0.417, se=0.175, p<0.017) all else being equal.

Infit

Model $\text{Infit}_A$ indicates that there was no significant variation in the log-odds of aberrant test response patterns on the Infit variable at the school level (Table 18). The results of $\text{Infit}_B$ indicate that there is a significant gender effect ($\beta$= -0.902, se=0.395, p=0.022). Female students are more frequently associated with a decrease in the occurrence in the log-odds of an aberrant response patterns. Model $\text{Infit}_C$ indicated a significant erasure behavior effect ($\beta$=0.197, se=0.074, p=0.005). Students who erased on the assessment more than average are more frequently associated with an increase in the occurrence in the log-odds of an aberrant response pattern, holding gender, economic status, and proficiency constant. Also model $\text{Infit}_C$, indicated a statistically significant difference in the log-odds of the occurrence of aberrant response patterns between the advanced proficient mathematics group and the proficient mathematics group ($\beta$=-1.856, se=0.761, p=0.015) all else being equal. There was a statistically significant difference in the log-odds of the occurrence of aberrant response patterns between the partially proficient mathematics group and the proficient mathematics group ($\beta$=0.820, se=0.375, p=0.029) all else being equal.

**Reading**

Outfit

The results of the unconditional model ($\text{Outfit}_A$) suggest there is significant variation ($\tau$=0.088, $\chi^2$=101.336, p<0.001) in the log-odds of an aberrant test response pattern on the Outfit variable at the school level (Table 19). The results of $Outfit_B$ indicate that there is a significant economic status effect ($\beta$=0.355, se=0.095, $p<0.001$). Economically disadvantaged students are more frequently associated with an increase in the occurrence in the log-odds of an aberrant pattern, holding constant gender. In model $Outfit_C$, there was a significant erasure behavior effect ($\beta$=0.092, se=0.035, p=0.008). Students who erased on the assessment more than average are more frequently associated with an increase in the occurrence in the log-odds of an aberrant response pattern, holding gender, economic status, and proficiency constant. There was a statistically significant difference in the log-odds of the occurrence of an aberrant response pattern between the advanced students and the proficient students ($\beta$=-0.432, se=0.201, p=0.031) all else being equal. There was also a statistically significant difference in the log-odds of the occurrence of an aberrant response pattern between the partially proficient students and the proficient students ($\beta$=0.404, se=0.091, $p<0.001$) all else being equal.

Infit

The results of the unconditional model ($Infit_A$) suggest there is significant variation ($\tau$=0.130, $\chi^2$=77.565, $p<0.05$) in the log-odds of an aberrant test response pattern on the Infit variable at the school level (Table 20). The results of $Infit_B$ indicate that there is a significant economic status effect ($\beta$= 0.858, se=0.131, $p<0.001$). Economically disadvantaged students are more frequently associated with an increase in the occurrence in the log-odds of an aberrant response patterns, holding gender constant. In model $Infit_C$, there was also a significant economic status effect ($\beta$= 0.394, se=0.136, p=0.004). Economically disadvantaged students are more frequently associated with an increase in the occurrence in the log-odds of an aberrant response patterns holding gender, erasure behavior, and proficiency constant. Model $Infit_C$ also indicated a significant erasure behavior effect (β=0.107, se=0.050, p=0.031). Students who erased on the assessment more than average are more frequently associated with an increase in the occurrence in the log-odds of an aberrant response pattern, holding gender, economic status, and proficiency constant. There was a statistically significant difference in the log-odds of the occurrence of aberrant response patterns between the advanced proficient students and the proficient students (β=-2.193, se=1.009, p<0.05) all else being equal. There was a statistically significant difference in the log-odds of the occurrence of aberrant response patterns between the partially proficient students and the proficient students (β=1.405, se=0.149, p<0.001) all else being equal.

## Discussion

In this case study I examined the effect of two demographic (gender and economic status) and two behavioral (erasure behavior and proficiency level) variables on person misfit within a multilevel (student and school) framework. It is hypothesized in this study that the individual and school level factors may make a significant contribution to variance in person fit indices. The data suggest that variations in person misfit across schools are statistically significant for the Outfit statistic for mathematics and reading and for the Infit statistic for reading (Research Question #1).

Investigations of the regression coefficients allowed for an analysis of the predictive ability of the demographic and behavioral variables on aberrant responding (Research Question 2). Analyses indicate that proficiency level is a statistically significant predictor of the likelihood of aberrant responding on both the mathematics and reading assessments. Highly proficient students exhibit higher levels of person misfit.

Through qualitative appraisals, Petridou and Williams (2007) found that similar findings in their study of mathematics ability and aberrant responding suggested that the significant effect on ability was explained by student carelessness affecting both Outfit and Infit results.

Gender is also a statistically significant predictor of aberrance on the mathematics assessment. It was found that female students were associated with decreases in the likelihood of aberrant responses. However, the effect of gender was eliminated with the addition of the behavioral variables, erasures and proficiency to the model suggesting that this effect was not due to gender but rather other factors.

Economic status and erasure behavior were found to be statistically significant on the reading assessment for both the Outfit and Infit statistics. It is likely that these variables are confounded with proficiency in their effect on students aberrant responding. The effect of erasures on the likelihood of aberrant responses was also found for the Infit model for mathematics assessment.

Although there is a long history of research on person fit, few researchers have directly examined person fit in a multilevel framework. This study extends research by Petridou and Williams (2007) by expanding the student level variables investigated to include economically disadvantaged and erasure behavior as wells as examining a new level of analysis, school-level. Results suggest that level of proficiency, gender, economic status, and erasure behavior are major correlates of person misfit.

# CHAPTER FIVE: DISCUSSION & SUMMARY

In this dissertation I have examined the usefulness of differential person functioning analysis as a method to assess invariant measurement. Within a Rasch framework, good model-data fit is necessary in order to have invariant measurement. Therefore it is essential that the model-data fit of assessments be examined to see how closely the requirements of invariant measurement are approximated. In particular, model-data fit must be determined to assess threats to the validity of person scores on an assessment. Differential person functioning (DPF) exists as a form of construct-irrelevant variance, skills or characteristics of the examinee that are not intended to be included in the measures (Ackerman, 1992). These ideas are explored in depth throughout the dissertation in the form of a comprehensive literature review and two case studies utilizing methods for assessing DPF within a high stakes assessment. In particular the following questions were used to guide the research:

(1) What is differential person functioning?

(2) How do the methods for assessing differential person functioning differ across contexts?

(3) To what extent does differential person functioning contribute to our understanding of person fit across contexts?

This chapter is divided into two sections, the first addresses each guiding question posed in the dissertation. The latter section identifies limitations to the study and implications for research, policy, and practice.

## Guiding Question #1: Differential Person Functioning

In this dissertation I present a historical depiction of the development of theories surrounding the investigation of differential person functioning (DPF) by researchers. Though classified by many other terms over the last 70 years, the ultimate concept has remained the same, DPF is a phenomena in which an individual's observed response pattern differs from the expected response pattern for individuals with the same location on the latent variable or construct. In other words, "the variability of the individual" impacts the reliability of the test score (Mosier, 1940, p. 357). In a time when policy decisions are being made based on aggregate test data it is important to remember Mosier's (1940) work on the variability of the individual and more current works which indicate that response patterns can still be unique for each person despite a groups common test score (Lumsden, 1977; Quaynor, Perkins, & Engelhard, 2009). Though not yet labeled as DPF in 1940 by Mosier or even 1977 by Lumsden, the idea that individuals can vary on their responses to an assessment, and that this variability influences the reliability of the assessment, represents the commencement of theories surrounding DPF.

Methods for assessing DPF have evolved with the development of new statistical and measurement procedures. Within Chapter Two I explored the two broad areas of analyses for DPF addressed by researchers. These include *Person Response Functions* and *Person Fit Indices*.

The person response function, analogous to the item response function, is being used by researchers as a way to gain a graphical representation of an individual's response pattern and subsequently contrast multiple individuals visually. When the PRFs of individuals cross, this is seen as a violation of the requirements of invariant measurement (Perkins & Engelhard, 2009). Specifically, a more able person must always have a better chance of success on an item than a less able person: *non-crossing PRFs* (Engelhard, 2013). Thus PRFs can provide vital information in identifying and exploring DPF---*unexpected response patterns*---on an assessment.

In addition to the use of graphical representations to assess DPF, one can also employ various statistical measures to quantify an individual's data to model fit through the use of person fit indices. Numerous person fit indices have been examined by researchers ranging from parametric to non-parametric and spanning IRT based statistics and those within Classical Test Theory. Several studies have been performed to determine the most useful person fit statistic but no consensus has been reached across researchers as to which index is the most optimal. Yet some researchers believe that factors regarding the data under investigation, and the theories used to model the data should be considered when choosing person fit indices.

Defining, identifying, and assessing DPF have each been addressed. Yet why do unexpected response patterns exist? A variety of response behaviors have been linked to construct irrelevant variance. While these behaviors include alignment error, guessing and carelessness of the individual as a few examples, recently the most common response behavior that researchers are concerned with is that of cheating in the form of student cheating and improper proctor assistance. Many of these behaviors can be inferred based on an investigation of a student's response pattern both visually and numerically. Given this information regarding the historical significance and contemporary issues of DPF, I chose to investigate two methods of assessing DPF. This was done through the use of two case studies. Both case studies utilize the same data set for analyses, high stakes mathematics and reading achievement data of third grade students. The first case study used a Many Facets Rasch Model and erasure analysis to provide baseline data on student and school level person fit and erasure behavior. The second case study uses a hierarchical generalized linear model to understand the influence of achievement level, erasure behavior, gender and economic status on person fit.

**Guiding Question #2: The Role of Context**

The second guiding question posed in the dissertation is, how do the methods for assessing differential person functioning differ across contexts? This question is explored using the two case studies presented in Chapters Three and Four, each replicated within two content areas (mathematics and reading) yielding a total of four contexts that are explored. Recall that in the first case study a Many Facets Rasch Model was employed to examine person fit within the area of erasure analysis. The second case study examined the relationship of aberrant responding and a set of covariates within a hierarchical generalized linear model using student and school levels of analysis. Here is a list of the four contexts represented in the case studies:

- Context 1 – Many Facets Rasch Model and Mathematics
- Context 2 – Many Facets Rasch Model and Reading
- Context 3 – Hierarchical generalized linear model and Mathematics
- Context 4 – Hierarchical generalized linear model and Reading

In the first case study the method of analysis included a Many Facets Rasch Model (MFRM) to generate person fit indices. The resulting person fit indices were aggregated by school and assessed in relation to erasure behavior which was also aggregated by school. Results across content area (mathematics and reading) did not have significant variation. However, there was some evidence to suggest that school membership did impact wrong-to-right (WTR) erasures and achievement. Lastly, person fit was not found to have a significant relationship with school achievement. This last point is expanded upon in the next section.

While the method differed, the focus on DPF remained the same in the second case study. The second case study uses a hierarchical generalized linear model (HGLM) to explore the relationships between gender, socioeconomic status, total erasures and proficiency level on the likelihood of student misfit at the individual and school levels of analysis. Results varied over the mathematics and reading content areas. In this case study, person fit was dichotomized as the outcome variable. Significant results were found based on the use of person fit as a dependent variable and the results differed between the two person fit statistics (Outfit and Infit) examined. More in terms of person fit is discussed in the next section which addresses the last guiding question.

Exploring DPF across the mathematics and reading assessments enabled me to assess whether the methods I chose to examine in the dissertation were sensitive to content area differences and whether differences based on the variables of interest exists across the content areas. As discussed, the method for assessing DPF provided differing results based on content area for the second case study (Context 3 and Context 4) and not the first case study (Case Study 1 and Case Study 2). Context 3 suggests that within the mathematics content area, gender, proficiency level and in some cases total erasures yielded a significant relationship with student patterns of misfit. This differed from Context 4 in that within the reading content area, socioeconomic status, total erasures and proficiency levels yielded a significant relationship with student patterns of misfit. These findings indicate that within the framework of person fit, the HGLM is useful in providing results sensitive to content area. They also clearly indicate that there are factors affecting student patterns of misfit that differ based on content area. Figure 20 further illustrates this finding.

While WTR erasures were observed to have a relationship with achievement in Context 1 and Context 2 the relationship was what was expected given the pre/post erasure design choice. Small increases in achievement were seen while large increases were absent from the results. Within Context 3 and Context 4 total erasures were chosen as a covariate, as the overall erasure behavior was of interest in the second case study not solely the WTR erasures. Interestingly, total erasures were found to have a significant relationship with student patterns of misfit in the reading content area. Students who erased on the assessment more than average are more frequently associated with an increase in student patterns of misfit.

In all four contexts the individual and school levels of analysis were taken into account. I observed significant findings at the school level across all contexts. This finding suggests that investigations at the school level are prudent when examining DPF. As we know, Petridou and Williams (2007) observed significant findings at the classroom level when assessing person fit. Given the organizational structure of the educational system these findings are not unexpected and highlight the need for multilevel analyses in educational research.

Ultimately what these methods do have in common when examined across context is that scores can have different meanings despite common indices and commonalities amongst groups of students.

## Guiding Question #3: Person Fit Across Contexts

The final guiding question in the dissertation is, to what extent does differential person functioning contribute to our understanding of person fit across contexts? Like the prior question, the present question is explored using the two case studies, each replicated within two content areas (mathematics and reading) yielding a total of four contexts:

- Context 1 – Many Facets Rasch Model and Mathematics
- Context 2 – Many Facets Rasch Model and Reading
- Context 3 – Hierarchical generalized linear model and Mathematics
- Context 4 – Hierarchical generalized linear model and Reading

Recall that within Context 1 and Context 2 the person fit statistics of Outfit and Infit were calculated using a Many Facets Rasch Model (MFRM). These indices were assessed with erasure analyses across schools. However as indicated in Chapter Three no significant relationship was found based on the analyses. Thus when considering the impact of person fit indices on the assessment of DPF in these contexts I found that person fit does not contribute to the our understanding of DPF within these contexts.

The second case study, which coincides with Context 3 and Context 4, calculated person fit statistics, Outfit and Infit, with a two facet Rasch model for dichotomous variables. Outfit and Infit were subsequently dichotomized to parse misfit versus non-misfit as outcome variables within a hierarchical generalized linear model (HGLM). Results indicate that various covariates have significant relationships with the person fit statistics. Little variation was found between the Outfit and Infit results for each content area. It was observed that for the mathematics section the variance component was significant when Outfit was the outcome variable as opposed to Infit. When assessing gender as a covariate for the mathematics section, the effect of gender on the model was eliminated with the addition of the behavioral variables when Outfit was the outcome variable. However, when Infit was the outcome variable in the same situation the relationship of gender with student misfit remained.

Overall it appears that the ability of person fit statistics to aid in the understanding of DPF is dependent upon the context that is analyzed.

**Limitations**

The case for generalizability can be of concern to some researchers when considering the use of case studies in this dissertation as such I recognize that statistical generalizations to populations are not capable in this research. However, the case studies in this dissertation are generalizable to other empirical investigations of a similar nature. Additionally the theoretical and analytical findings from this dissertation are also generalizable (Yinn, 2009). The use of case studies in this dissertation means that the results of this study are not generalizeable to all students or all schools. In addition the data were obtained from a secondary source limiting my ability to choose variables and levels of analysis. I attempted to address this concern by ensuring the integrity of the data through rigorous data checks.

I did not intend to address all methods for investigating differential person functioning in this study. However, my intent was to highlight two key methods for assessing DPF in one population of students and schools. This study provides the reader with examples of the vastness of DPF and other methods for examining DPF may yield different results. The methods highlighted in this study demonstrate the complexity and density of DPF.

Lastly, data were not available for district level or teacher level units of analyses. Future research should include these variables in order to bring greater depth to this study. The dissertation does however explore the student and school levels and these units of analyses have generated useful findings.

## Implications for Research, Policy, and Practice

In this section, I discuss the importance of these findings for research in the area of differential person functioning. I then address implications for policy and practice in the area of student achievement. Lastly, I pose future directions for research of invariant measurement.

### Research

The duality between the commonly assessed threats to validity, differential item functioning and the less commonly addressed, differential person functioning allows for an ease of understanding regarding the extreme importance of examining both threats when conducting validity analyses. As an explorative study, this dissertation enables readers to begin a discussion about the importance of such analyses and theoretical ideas. In this dissertation, I highlight the connection between differential person functioning and person fit analysis and argue that a combination of methods is useful in examining invariant measurement within an assessment. The goal in assessment development is invariant measurement in item calibration, person measurement and a common attribute continuum. This dissertation takes a closer look at the area of person measurement. Recall the requirements for invariant measurement:

*Item calibration:*

1. The calibration of the items must be independent of the particular persons used for calibration: *Person-invariant calibration of test items.*
2. Any person must have a better chance of success on an easy item than on a more difficult item: *Non-crossing item response functions.*

*Person measurement:*

3. The measurement of persons must be independent of the particular items that happen to be used for the measuring: *Item-invariant measurement of persons.*
4. A more able person must always have a better chance of success on an item than a less able person: *Non-crossing person response functions.*

*Variable map:*

5. Person and items must be located on a single underlying latent variable: *Unidimensionality.*

My goal was to provide two clear yet different examples of studies that can assess differential person functioning within a common data set. This was accomplished using person fit analysis as the common global method of analyses and Rasch measurement as the common theoretical framework. Then within each case study additional differing methodologies were employed to address sets of research questions. Upon completion of these analyses I have chronicled some lessons learned as they relate to invariant measurement and Rasch measurement theory when concerned with DPF.

- This study is the first to include erasure behavior as a student variable of interest combining two extremely important areas of research, erasure analyses and student misfit.

- This research provided an empirical example of the importance of assessing student variability as it relates to the reliability of assessment scores.

- The choice of method used to assess DPF is important. Thus the use of multiple methods has proved useful in providing a variety of important and compelling information.

**Policy and Practice**

This study has practical implications for teachers, test developers, and policy makers. Understanding the substantive impact of statistically significant person misfit patterns can allow educators to explore what this means for the instructional needs of each student. Specifically this study found that the school a student attends matters. For example, patterns of wrong to right erasures, which we know at high levels can be telling of improper behaviors, were found to be consistently different for certain schools in the study. Also when considering a hierarchical generalized linear model in the context of person fit a significant amount of variance in aberrant responding was accounted for at the school level.

In terms of the covariates examined in the second case study, proficiency level, gender, economic status, and erasure behavior, several implications can be drawn from the findings. As it relates to proficiency level and economic status, significant findings on the relationship of these variables and aberrant responding could likely mean that the assessment was not properly aligned with the student's knowledge level resulting in a mis-measurement of the construct for these students. While the policy implications for the significant findings related to gender as a covariate can suggest that policy be restricted such that all students (male and female) are obtaining beneficial support, particularly in the area of mathematics. Results of this study indicate that while gender was a significant predictor of the likelihood aberrant responding that this effect was accounted for by other covariates that were investigated. Opportunity to learn is the common thread that weaves together the preceding findings. Does every student have access to the resources they need to be successful? Have the students been exposed to the curriculum presented to them within the assessment? Finally, findings suggest that further investigation be undertaken to examine the types of erasures students perform and the choices behind such decisions such as carelessness, instruction directions, and improper assistance.

Lastly, I believe systematic methods for assessing student aberrance within a multilevel framework should be added to routine analyses performed by test developers, school districts and state boards of education as a way to easily and regularly provide educators and policy makers with critical information.

**Future research**

While this dissertation offers new ideas of DPF from a quantitative perspective, qualitative appraisals of DPF are capable of providing a level of understanding that one cannot draw from a quantitative analysis. A comprehensive mixed methods approach to DPF is needed. This research can build on Petridou and Williams (2010) and their connecting of both quantitative and qualitative work (Petridou & Williams, 2007). Additional research in the area of mathematics achievement and person fit in a student and classroom model combined with qualitative investigations can provide a deeper understanding of DPF.

In this dissertation, I examine the role of the school context in this type of research, Petridou and Williams (2007) examined the role of classrooms in similar research, more research is needed to investigate the teacher and district levels of analysis on this work. This explorative study uses empirical data making generalizations difficult, especially as it relates to Outfit and Infit whose values are dependent on the data. Thus simulation studies could be useful in gaining insight into the ability to generalize the ideas explored in this study. As well applications are needed across grades, content areas, and geographical locations of students.

Overall, this work supports the use of DPF as a methodological approach for examining the validity of each person's response pattern. This level of detail is needed in order to add to confidence in the appropriateness of the scores assigned to each person, and the decisions that are made on the basis of these scores. It is clear that additional research is needed in the area, but the results of this dissertation highlight the potential benefit of adding these analyses to the routine data checking processes used in educational assessments.

**References**

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

Amerin-Beardsley A., Berliner D. C., Rideau S. (2010). Cheating in the first, second, and third degree: Educator's responses to high-stakes testing. *Educational Policy Analysis Archives, 18*(14).

Armstrong, R. D., Stoumbous, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the $l_z$ person-fit statistic. *Practical Assessment, Research & Evaluation, 12*(16), Retrieved on September 18, 2012 from http://pareonline.net/pdf/v12n16.pdf.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

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

Carroll, J. B., Meade, 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 a social science* (pp. 109-143). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chatman, S. P. (March 1985). *The relationship between response pattern aberrance and course performance in math placement.* Paper presented at the American Educational Research Association Annual Meeting, Chicago, IL.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice. 17*(1), 33-44.

Cizek, G. J. (2003). When teachers cheat. *Education Digest, 68*(6), 28-31.

Conijn, J. M., Emons, W., van Assen, M., & 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.

Dodden, H. & Darabi, M. (2009). Person-fit**:** Relationship with four personality tests in mathematics. *Research Papers in Education*, *24*(1), 115-126.

Donlan, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group determined item difficulties. *Educational and Psychological Measurement, 28,* 105–113.

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Engelhard, G. (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,* 585-602.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer et al. *Measurement and prediction*. *The American soldier* Vol 4. New York, NY: Wiley.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139-150. Retrieved September 3, 2009, from SocINDEX with Full Text database.

Hambleton R. K. (1989). Principles and selected applications of item response theory. In Linn L.R. (Ed.), *Educational measurement* (pp. 147-200). New York: Macmillan.

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionably test data and dissimilar item difficulties. *Educational and Psychological Measurement*, *18*(3), 133-146.

Harnisch, D. L. & Tatsuoka, K. K. (1983). A comparison of appropriateness indices based on item response theory. In R. K. Hambleton (Ed.), *Applications of Item Response Theory* (pp 105-123). Vancouver, BC: Educational Research Institute of British Columbia.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow Jones-Irwin

Johanson, G. & Alsmadi, A. (2002). Differential person functioning. *Educational and Psychological Measurement, 62*(3), 435-443.

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

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298.

Keats, J. A. (1967). Test theory. *Annual Review of Psychology*. *18*, 217-238.

Klauer, K. C. & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical & Statistical Psychology*. *43*(2), 193-206.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, *4*, 269-290.

Li, M. F. & Olejnik, S. (1997). The power of Rasch person fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*,215-231.

Linacre, J. M. (2010). FACETS (version 3.66.2) [Computer program]. Chicago: MESA.

Linacre, J. M. (2009). *A user's guide to FACETS Rasch-model computer programs*. Chicago, IL: MESA Press.

Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement*, *4*(1), 1-7.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement*, *1*(4), 477-482.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*(1), 15-29.

Mead, R., Anderson, K., Korts, J., (April, 2011) *Erasures and Rasch residuals*. Paper Presented at National Council on Measurement in Education Annual Meeting, New Orleans, LA.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*, 3-8.

Meijer, R. R., Muijtjens, A. M. M., & van der Vleuten, C. P. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education*, *9*(1), 77-89.

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

Miller, D. M. (1986). Time allocation and patterns of item response. *Journal of Educational Measurement, 23*(2), 147-156.

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

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

Perkins, A. (2010). *Examining Differential Person Functioning on Mathematics Items Related to Home Language, Social Class, and Gender within France, Germany, Hong Kong, and the United States* (unpublished empirical study). Emory University, Atlanta, GA.

Perkins, A., & Engelhard, G. (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*(3), 227-247.

Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practice, 20*(1).

Quaynor, L., Perkins, A., & Engelhard, G. (October, 2009). *Differential item and person functioning related to home language on the PISA 2003 mathematics test.* Paper presented at the 2009 Georgia Educational Research Association Symposium.

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

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.).Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A., Cheong, Y. F., Congdon, R., & Du Toit, M. (2011). *HLM7: Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.

Reise, S. P. (1990). A comparison of item-and person fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*(2), 127-137.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20,* 207-219.

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

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

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden*, 7, 131-145.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66,*191-207.

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

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

Smith, R. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, *46,* 359-372.

Smith, R. M. & Hedges, L.V. (1982). A comparison of likelihood ratio $\chi 2$ and Pearsonian $\chi 2$ tests of fit in the Rasch model. *Educational Research and Perspectives,* 9 44-54.

Tatsuoka, K. K., & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20,* 221–230.

Trabin, T. E. & Weiss, D.J. (1983). The person response curve: Fit of individuals to item response theory models. In D.J. Weiss (Ed.). *New horizons in testing: Latent trait test theory and computer adaptive testing* (pp.83-108). New York: Academic Press.

Trabin, T. E. & Weiss, D. J. (1979). *The person response curve: Fit of individuals to item characteristic curve models.* (Research Report 79-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Vale, C.D. & Weiss, D. J. (1975). *A study of computer-administered stradaptive testing.* (Research Report 75-4, NTIS No. Ad-A018758) Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*(3), 267-298.

van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics*, *37*.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test*. (Research Report 73-3, NTIS No. AD-768376). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wright, B. D. (1977). Solving the measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97-116.

Wright, B.D. (1967). Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems* (pp. 85-101). Princeton, NJ: Educational Testing Service.

Wright, B.D., Linacre, J.M., Gustafson, J.E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions,* 8(3), 370. Retrieved September 24, *2002* from http://rasch.org/rmt/rmt83b.htm.

Wright, B.D., & Stone, M. (1979). *Best test design: Rasch measurement*. Chicago, IL: MESA Press.

Yinn, R. K. (2009). *Case study research. Design and methods* (4th ed.). Thousand Oaks, CA: Sage Publications.

Zumbo, B. D. (1999). *Handbook of the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, National Defense Headquarters.

**Table 1. Chronological List of Key Ideas in Person Measurement**

| Author(s) | Year | Theme | Key Ideas |
|---|---|---|---|
| Mosier | 1940/41 | RPM | • Introduces the idea of person and item invariance in the area of psychophysics.<br>• Emphasizes the necessity of taking "into account the variability of the individual" with respect to a set of items (p. 356).<br>• Mosier identifies that the reliability of a set of items may vary from one individual to the next. |
| Keats | 1967 | RPM, PRF | • Examines proposed solutions to eliminating the interference of individual characteristics on item responses.<br>• Subjects should be grouped based on overall test score rather than observed individually.<br>• For persons to be ordered "each item would have to discriminate significantly between groups" (p. 218).<br>• Keats key concern is "whether or not subjects have been ordered with respect to more than one dimension" (p. 218). |
| Weiss | 1973 | RPM, PRF | • Proposed the notion of a person "trace line", a graphical representation of item difficulties and individual responses to items. As item difficulty increased, the individual's percentage correct would decrease. |
| Vale & Weiss | 1975 | RPM, PRF | • Investigated the premise that more consistent individuals would yield more stable ability estimates.<br>• Studied the test-retest reliability of trace line plots introducing the concept of subject characteristic curves. |
| Levine & Rubin | 1976 | PFI | • Explore student response patterns to multiple choice items on the Scholastic Aptitude Test<br>• Develop appropriateness indices which are measures of goodness of fit of psychometric models to individual's response patterns.<br>• Levine and Rubin see these indices as useful measures for identifying spuriously high and spuriously low examinees. |

**Table 1 (continued). Chronological List of Key Ideas in Person Measurement**

| Author(s) | Year | Theme | Key Ideas |
|---|---|---|---|
| Lumsden | 1977/80 | RPM, PRF | • Using psychological measurement and mental growth as the backdrop for his work, Lumsden presents a useful approach to addressing issues of person reliability.<br>• Introduces the person characteristic curve (PCC): “The person characteristic curve is the plot for a single subject of the proportion of items passed at different difficulty levels. It is perfectly analogous to the item characteristic curve.” (p. 478)<br>• Lumsden identifies issues of using total test score for grouping individuals when addressing issues of ordering persons.<br>• Examines situations of crossing PCC’s in which two subjects receive the same total test score but when they are examined in relation to their correct responses on items ordered by difficulty their PCC’s cross.<br>• Crossing of PCCs will result in the estimates of reliability being “biased by the difficulty of the items” (p. 481). |
| Trabin & Weiss | 1979 | RB, PRF | • Compared expected and observed PRCs identify that 90% of students fit the expected PRCs.<br>• Study indicated that expected PRCs served as a good predictor of observed PRCs.<br>• Introduced the idea of a student “profile”. |

**Table 1 (continued). Chronological List of Key Ideas in Person Measurement**

| **Author(s)** | **Year** | **Theme** | **Key Ideas** |
|---|---|---|---|
| Van der Flier | 1982 | PFI | • Presents a study on deviant scores through the lens of cross-cultural psychology. Asserting that there are many problems when comparing the test scores of groups from differing cultural backgrounds.<br>• Notes that if a test has some given meaning at a group level comparison it doesn't indicate that the same meaning holds at the individual level.<br>• In this study, deviance scores are observed in more detail. Deviance scores represent how much an individual's observed score pattern differs from an expected score pattern.<br>• The meanings of deviance scores are specific to the test and population of interest. Generalizations of meanings to other contexts are problematic. |
| Rudner | 1983 | PFI | • Concerned with the systematic investigation of individual performance on assessments.<br>• Evaluates nine proposed indices for assessing person fit based on Rasch (1960) and Birnbaum (1968) models as well as correlation coefficients and item sequencing.<br>• Assessed that the power of the application would be a possible criteria reason for selecting one proposed index over another. Other criteria included available item parameters and computation requirements. Additionally the type of assessment might dictate the selection of the critical value.<br>• Large scale as opposed to classroom tests might be better suited to employ a conservative decision rule. |

**Table 1 (continued). Chronological List of Key Ideas in Person Measurement**

| Author(s) | Year | Theme | Key Ideas |
|---|---|---|---|
| Masters | 1988 | PFI | • Provides an investigation of the traditional discrimination parameter found in the 2PL and 3PL IRT models.<br>• Postulates that high discriminations, normally a desirable characteristic of an item, could be an indication of a greater measurement issue. Highly discriminating items effectively discern between high achieving and low achieving students.<br>• Suggest that the Rasch model is unique in that it proposes items with curiously high discriminations be eliminated from the test. The Rasch model motivates researchers to ask why an item discriminates well.<br>• Suggest that this differential item performance could be the result of a number of issues including opportunity to answer and test wiseness. |
| Klauer & Rettig | 1990 | PFI | • Propose a standardized person test for assessing consistency with a latent trait model.<br>• Compare three statistics as options for evaluating the invariance hypothesis. The statistics are computed using single response vectors that are standardized such that their conditional probabilities do not depend upon the absolute value of an individual's theta. |

**Table 1 (continued). Chronological List of Key Ideas in Person Measurement**

| Author(s) | Year | Theme | Key Ideas |
|---|---|---|---|
| Reise | 1990 | PFI | • Compares a $\chi 2$ item-fit index with a likelihood-based person-fit index.<br>• Demonstrates that a traditional item-fit index can be used to assess person fit (and vice versa).<br>• Exhibits that while many item fit indices work well to identify misfitting items the same indices were not as successful with identifying severely misfitting individuals.<br>• Recommended a likelihood-based index in applied fit analyses to evaluate both examinee and item mode-data fit. |
| Meijer | 1996 | RB | • Provide an overview of person fit research.<br>• Person fit research was initially an explorative endeavor but is morphing into a method used for providing a stronger case to support a suspected deviance behavior. |
| Sijtsma & Meijer | 2001 | PRF, PFI | • Outlines the invariance requirements for PRFs using a nonparametric IRT framework.<br>• Defines PRFs in detail while juxtaposing its characteristics with that of IRFs.<br>• Explores the use of PRFs as a method for identifying aberrant responses as opposed to the use of person fit statistics.<br>• Discuss the problematic nature of the 2PL, 3PL, and 4PL models when defining a PRF because with these models PRFs have the ability to intersect.<br>• The PRF is a function with the potential to diagnosis aberrance. |

**Table 1 (continued). Chronological List of Key Ideas in Person Measurement**

| **Author(s)** | **Year** | **Theme** | **Key Ideas** |
|---|---|---|---|
| Embretson & Reise | 2000 | PRF, PFI | • Sketches out problems that scholars have identified within the area of person fit research such as the inability to assess the causes for the unexpected behaviors detected with person fit indices.<br>• Embretson and Reise point out the importance of conducting validity studies to demonstrate the importance of real life situations as appraisals of student knowledge as opposed to the use of test scores.<br>• Suggests the linkage of student achievement to construct validity through the use of PCCs.<br>• States that person fit statistics should be used in conjunction with PCC interpretations in order to fully evaluate student response patterns. |
| Karabatsos | 2003 | RB, PFI | • Presents a comprehensive analysis comparing various parametric and non-parametric fit statistics under differing test conditions to ascertain a decision as to which person fit statistics is best suited and the virtues of each statistic. |
| Engelhard | 2008 | PFI | • Discusses DIF and DPF as a lack of mode-data fit.<br>• Emphasizes the observation that the presence of DIF and DPF signify that the requirements of invariant measurement were not met by the item-person responses.<br>• Study uses residuals, standardized residuals, and mean square error statistics (Outfit) to examine person fit in the assessment of students with disabilities.<br>• Proposes the development of a mixed methods approach to psychometrics. |

*Note.* RPM = Reliability of Person Measurements, RB = Response Behavior, PRF= Person Response Functions, PFI = Person Fit Indices

**Table 2. Abbreviated List of Person Fit Indices**

| Description | Equation | Parameters |
|---|---|---|
| *Personal Point biserial correlation ($rb_i$)*<br>Donlon and Fisher (1968) | | |
| • Differs from $r_i$ in that it assumes that the underlying variable is normally distributed. | $r_{bi} = \text{Corr}(X_n, p)$ | $X_n$, examinee *n*'s scored item response vector<br><br>$P$, item vector of proportion correct |
| *Appropriateness Indices*<br>Levine & Rubin (1979) | | |
| • Measure of fit of a psychometric model to an individual's item response set. The authors consider this index as a broad low power identifier of irregular student response patterns.<br>• Levine and Rubin saw these indices as useful measures for identifying spuriously high and spuriously low examinees.<br>• It is important to note that this measure is only a function of the examinee's responses as such appropriateness indices are often closely related to total test score.<br>• Appropriate indices are further discussed in the work of Harnisch and Tatsuoka (1983). | $l = \sum_{j=1}^{J} [X_{nj}(\ell n\, P_{nj1}) + (1 - X_{nj})(\ell n\, P_{nj0})]$ | $J$, number of items<br><br>$X_{nj}$, examinee's scored response to test itme $j$<br><br>$P_{nj1}$, probability of a correct ($X_{nj} = 1$) response<br><br>$P_{nj0}$, probability of an incorrect ($X_{nj} =$ 0) response, with $P_{nj0} = (1- P_{nj1}$ ). |

**Table 2 (continued). Abbreviated List of Person Fit Indices**

| **Description** | **Equation** | **Parameters** |
|---|---|---|
| *Norm Conformity Index ($NCI_i$)*<br>Tatsuoka and Tatsuoka (1982) | | |
| • Mathematically related to van der Flier's (1982) deviance score | $NCI_i = 2\ Sa/S - 1$ | *Sa*, sum of the above diagonal elements of the dominance matrix from the ordered item response vector<br><br>*S*, sum of all matrix elements |
| *Modified Caution Index ($C_i$)*<br>Rudner (1983) | | |
| • Based on caution indices originally introduced by Sato in 1975 (as cited in Harnisch & Linn, 1981) and Harnisch & Linn (1981). Found to be a very stable measure of fit. | $C_i = \dfrac{\sum_{j=1}^{n_{i.}} (1 - u_{ij})\, n_{.j} - \sum_{j=n_{i.}+1}^{N} u_{ij}\, n_{.j}}{\sum_{j=1}^{n_{i.}} n_{.j} - \sum_{j=N+1-n_{i.}}^{N} n_{.j}}$ | $u_{ij}$, observed item response<br><br>$n_i$, the total score for examinee *i*<br><br>$n_j$, the number of correct responses to item *j* |

**Table 2 (continued). Abbreviated List of Person Fit Indices**

| Description | Equation | Parameters |
|---|---|---|
| *$H^T$*<br>As discussed in Karabatsos (2003) | | |
| • Of a comprehensive analysis of 36 fit indices, it was determined that the $H^T$ statistic was not only the best for identifying students with unusual response patterns generally but also with detecting aberrant examinees on exams of varying lengths and examinees with a variety of different response behaviors. | $H^T = \frac{\sum_{n \neq m} \beta_{nm} - \beta_n \beta_m}{\sum_{n \neq m} \min \{\beta_n(1 - \beta_m), (1 - \beta_n)\beta_m\}}$ | $\beta_{nm}$, the covariance between the scored test responses of examinee n with examinee m, with $\beta_{nm} = J^{-1} \sum_{j=1}^{J} X_{nj} X_{mj}$<br><br>$\beta_n$, proportion correct for examinee n over the J test items, $\beta_n = J^{-1} r_n$ |
| *likelihood-based person-fit index*<br>Reise (1990) | | |
| • Reise found that the Likelihood based person fit index ($Z_3$) was more efficient than $\chi 2$. In particular, he found that $Z_3$ was able to identify two types of misfitting behavior, "(1) response vectors that are less consistent than the model predicts, and (2) response vectors that are too consistent with respect to the specified model" (p. 135). | $L\|\theta = \sum_{k=1}^{K} \{U_k [\ln P_k(\theta)] + (1 - U_k) \times [\ln Q_k(\theta)]\}$ , | *k*, number of items<br><br>$U_k$, 0,1 item response<br><br>$P_k(\theta)$, probability of a correct response given θ<br><br>$Q_k(\theta)$, 1- $P_k(\theta)$ |

**Table 3. Summary of Case Studies**

| Purpose | Research Questions | Methodology | Results |
|---|---|---|---|
| *CASE STUDY ONE: Using Person Fit to Examine Erasure Data* | | | |
| To explore the relationship between wrong-to-right erasures, person fit indices, and school-level mathematics and reading achievement using a pre/post erasure design. | 1. Is there a relationship between wrong-to-right erasures and mathematics and reading achievement at the school level?<br>2. Does the relationship between wrong-to-right erasures and mathematics and reading achievement vary based on school context?<br>3. Is person fit a useful index for detecting irregular erasure behavior at the school level? | • Pre/Post Erasure Design<br>• Many Facets Rasch model | • Person fit indices identified misfitting students; however, there were no systematic patterns to provide explanations for variations in person fit. Analyses did identify two schools with unexpected increases in achievement based on erasure analyses. |
| *CASE STUDY TWO: Using a Multilevel Model to Examine Person Fit* | | | |
| To examine student and school factors that may be associated with the aberrant responses of students that include mathematics proficiency levels, economic status, gender, and student erasure behavior. | 1. What proportion of aberrant responding is attributable to student- and school-level factors?<br>2. Do select student- and school- level factors predict aberrant responding? | • Hierarchical generalized linear modeling with dichotomous dependent variables | • Gender and achievement are significant predictors of aberrant responses on the mathematics assessment.<br>• Economic status, erasure behavior, and proficiency are significant predictors of aberrant responses on the reading assessment |

*Note.* Each application will be repeated in the content areas of mathematics and reading.

**Table 4. Student Demographics for Grade 3 students in Case Study**

| | | N=4,248 | % |
|---|---|---|---|
| **Gender** | | | |
| | Male | 2,195 | 51.67 |
| | Female | 2,050 | 48.26 |
| | Missing | 3 | 0.07 |
| **Ethnicity** | | | |
| | White | 2,359 | 55.53 |
| | Asian | 813 | 19.14 |
| | Hispanic | 576 | 13.56 |
| | African-American | 470 | 11.06 |
| | Pacific Islander | 3 | 0.07 |
| | American Indian | 5 | 0.12 |
| | Unknown | 22 | 0.52 |
| **Economically Status** | | | |
| | No | 3,022 | 71.14 |
| | Yes | 1,226 | 28.86 |
| **Math Proficiency Level** | | | |
| | Advanced Proficient | 1,695 | 39.90 |
| | Proficient | 1,726 | 40.63 |
| | Partially Proficient | 772 | 18.17 |
| | Missing | 55 | 1.29 |
| **LAL Proficiency Level** | | | |
| | Advanced Proficient | 309 | 7.27 |
| | Proficient | 2,378 | 55.98 |
| | Partially Proficient | 1,499 | 35.29 |
| | Missing | 62 | 1.46 |
| **Mean Math Scale Score (SD)** | | 236.41 (41.08) | |
| **Mean LAL Scale Score (SD)** | | 207.09 (26.58) | |

**Table 5. Student Erasures by School (Mathematics Content Area)**

| | N | Total Erasures | % WR | % RW | % WW | Mean WR per student | Mean RW per student | Mean WW per student |
|---|---|---|---|---|---|---|---|---|
| **901** | 57 | 101 | 62.38 | 14.85 | 22.77 | 1.11 | 0.26 | 0.40 |
| **902** | 95 | 164 | 71.34 | 10.98 | 17.68 | 1.23 | 0.19 | 0.31 |
| **903** | 16 | 49 | 89.80 | 0.00 | 10.20 | 2.75 | 0.00 | 0.31 |
| **904** | 68 | 101 | 76.24 | 9.90 | 13.86 | 1.13 | 0.15 | 0.21 |
| **905** | 81 | 75 | 64.00 | 17.33 | 18.67 | 0.59 | 0.16 | 0.17 |
| **906** | 84 | 117 | 64.96 | 16.24 | 18.80 | 0.90 | 0.23 | 0.26 |
| **907** | 99 | 142 | 55.63 | 14.08 | 30.28 | 0.80 | 0.20 | 0.43 |
| **908** | 46 | 215 | 74.42 | 4.65 | 20.93 | 3.48 | 0.22 | 0.98 |
| **909** | 73 | 135 | 71.85 | 14.07 | 14.07 | 1.33 | 0.26 | 0.26 |
| **910** | 89 | 82 | 73.17 | 12.20 | 14.63 | 0.67 | 0.11 | 0.13 |
| **911** | 83 | 138 | 57.97 | 11.59 | 30.43 | 0.96 | 0.19 | 0.51 |
| **912** | 158 | 238 | 56.72 | 13.03 | 30.25 | 0.85 | 0.20 | 0.46 |
| **913** | 44 | 84 | 65.48 | 17.86 | 16.67 | 1.25 | 0.34 | 0.32 |
| **914** | 62 | 113 | 41.59 | 22.12 | 36.28 | 0.76 | 0.40 | 0.66 |
| **915** | 42 | 77 | 67.53 | 12.99 | 19.48 | 1.24 | 0.24 | 0.36 |
| **916** | 181 | 161 | 71.43 | 11.80 | 16.77 | 0.64 | 0.10 | 0.15 |
| **917** | 79 | 114 | 64.91 | 14.91 | 20.18 | 0.94 | 0.22 | 0.29 |
| **918** | 98 | 100 | 64.00 | 21.00 | 15.00 | 0.65 | 0.21 | 0.15 |
| **919** | 119 | 173 | 61.27 | 15.61 | 23.12 | 0.89 | 0.23 | 0.34 |
| **920** | 60 | 91 | 60.44 | 17.58 | 21.98 | 0.92 | 0.27 | 0.33 |
| **921** | 102 | 259 | 68.73 | 10.04 | 21.24 | 1.75 | 0.25 | 0.54 |
| **922** | 9 | 14 | 85.71 | 7.14 | 7.14 | 1.33 | 0.11 | 0.11 |
| **923** | 109 | 134 | 63.43 | 10.45 | 26.12 | 0.78 | 0.13 | 0.32 |
| **924** | 122 | 189 | 61.38 | 17.46 | 21.16 | 0.95 | 0.27 | 0.33 |
| **925** | 99 | 176 | 68.75 | 10.80 | 20.45 | 1.22 | 0.19 | 0.36 |
| **926** | 88 | 168 | 63.69 | 16.07 | 20.24 | 1.22 | 0.31 | 0.39 |
| **927** | 26 | 71 | 57.75 | 23.94 | 18.31 | 1.58 | 0.65 | 0.50 |
| **928** | 91 | 120 | 70.83 | 14.17 | 15.00 | 0.93 | 0.19 | 0.20 |
| **929** | 166 | 248 | 65.32 | 12.10 | 22.58 | 0.98 | 0.18 | 0.34 |
| **930** | 135 | 165 | 64.85 | 13.33 | 21.82 | 0.79 | 0.16 | 0.27 |
| **931** | 116 | 240 | 72.92 | 8.75 | 18.33 | 1.51 | 0.18 | 0.38 |
| **932** | 157 | 239 | 66.11 | 15.48 | 18.41 | 1.01 | 0.24 | 0.28 |
| **933** | 67 | 68 | 73.53 | 8.82 | 17.65 | 0.75 | 0.09 | 0.18 |
| **934** | 92 | 126 | 57.14 | 14.29 | 28.57 | 0.78 | 0.20 | 0.39 |
| **935** | 78 | 156 | 60.26 | 13.46 | 26.28 | 1.21 | 0.27 | 0.53 |
| **936** | 108 | 173 | 60.12 | 17.34 | 22.54 | 0.96 | 0.28 | 0.36 |
| **937** | 42 | 71 | 73.24 | 11.27 | 15.49 | 1.24 | 0.19 | 0.26 |
| **938** | 80 | 95 | 74.74 | 13.68 | 11.58 | 0.89 | 0.16 | 0.14 |
| **939** | 102 | 194 | 69.59 | 12.37 | 18.04 | 1.32 | 0.24 | 0.34 |

**Table 5 cont. Student Erasures by School (Mathematics Content Area)**

| | N | Total Erasures | % WR | % RW | % WW | Mean WR per student | Mean RW per student | Mean WW per student |
|---|---|---|---|---|---|---|---|---|
| **940** | 89 | 118 | 67.80 | 14.41 | 17.80 | 0.90 | 0.19 | 0.24 |
| **941** | 85 | 216 | 61.57 | 12.96 | 25.46 | 1.56 | 0.33 | 0.65 |
| **942** | 115 | 165 | 75.76 | 9.09 | 15.15 | 1.09 | 0.13 | 0.22 |
| **943** | 61 | 89 | 73.03 | 13.48 | 13.48 | 1.07 | 0.20 | 0.20 |
| **944** | 20 | 29 | 72.41 | 10.34 | 17.24 | 1.05 | 0.15 | 0.25 |
| **945** | 154 | 259 | 67.95 | 11.20 | 20.85 | 1.14 | 0.19 | 0.35 |
| **946** | 87 | 218 | 63.76 | 12.84 | 23.39 | 1.60 | 0.32 | 0.59 |
| **947** | 64 | 94 | 64.89 | 10.64 | 24.47 | 0.95 | 0.16 | 0.36 |
| **948** | 150 | 127 | 78.74 | 9.45 | 11.81 | 0.67 | 0.08 | 0.10 |
| **Total** | 4248 | 6691 | 66.19 | 12.99 | 20.82 | 1.04 | 0.20 | 0.33 |

***Note**. WR: wrong to right, RW: right to wrong, and WW: wrong to wrong There are 35 multiple choice items in the mathematics content area. This differs from the reading content area which contains 18 multiple choice items.

**Table 6. Student Erasures by School (Reading Content Area)**

| | N | Total Erasures | % WR | % RW | % WW | Mean WR per student | Mean RW per student | Mean WW per student |
|---|---|---|---|---|---|---|---|---|
| **901** | 57 | 29 | 72.41 | 6.90 | 20.69 | 0.37 | 0.04 | 0.11 |
| **902** | 95 | 102 | 76.47 | 4.90 | 18.63 | 0.82 | 0.05 | 0.20 |
| **903** | 16 | 13 | 61.54 | 15.38 | 23.08 | 0.50 | 0.13 | 0.19 |
| **904** | 68 | 29 | 44.83 | 24.14 | 31.03 | 0.19 | 0.10 | 0.13 |
| **905** | 81 | 63 | 46.03 | 28.57 | 25.40 | 0.36 | 0.22 | 0.20 |
| **906** | 84 | 59 | 40.68 | 22.03 | 37.29 | 0.29 | 0.15 | 0.26 |
| **907** | 99 | 67 | 41.79 | 20.90 | 37.31 | 0.28 | 0.14 | 0.25 |
| **908** | 46 | 87 | 59.77 | 14.94 | 25.29 | 1.13 | 0.28 | 0.48 |
| **909** | 73 | 33 | 60.61 | 15.15 | 24.24 | 0.27 | 0.07 | 0.11 |
| **910** | 89 | 45 | 55.56 | 20.00 | 24.44 | 0.28 | 0.10 | 0.12 |
| **911** | 83 | 78 | 55.13 | 11.54 | 33.33 | 0.52 | 0.11 | 0.31 |
| **912** | 158 | 88 | 56.82 | 25.00 | 18.18 | 0.32 | 0.14 | 0.10 |
| **913** | 44 | 41 | 48.78 | 19.51 | 31.71 | 0.45 | 0.18 | 0.30 |
| **914** | 62 | 35 | 57.14 | 20.00 | 22.86 | 0.32 | 0.11 | 0.13 |
| **915** | 42 | 14 | 71.43 | 7.14 | 21.43 | 0.24 | 0.02 | 0.07 |
| **916** | 181 | 111 | 52.25 | 29.73 | 18.02 | 0.32 | 0.18 | 0.11 |
| **917** | 79 | 63 | 65.08 | 17.46 | 17.46 | 0.52 | 0.14 | 0.14 |
| **918** | 98 | 45 | 55.56 | 13.33 | 31.11 | 0.26 | 0.06 | 0.14 |
| **919** | 119 | 56 | 46.43 | 23.21 | 30.36 | 0.22 | 0.11 | 0.14 |
| **920** | 60 | 27 | 66.67 | 14.81 | 18.52 | 0.30 | 0.07 | 0.08 |
| **921** | 102 | 56 | 62.50 | 7.14 | 30.36 | 0.34 | 0.04 | 0.17 |
| **922** | 9 | 2 | 100.00 | 0.00 | 0.00 | 0.22 | 0.00 | 0.00 |
| **923** | 109 | 61 | 47.54 | 27.87 | 24.59 | 0.27 | 0.16 | 0.14 |
| **924** | 122 | 100 | 46.00 | 26.00 | 28.00 | 0.38 | 0.21 | 0.23 |
| **925** | 99 | 70 | 42.86 | 27.14 | 30.00 | 0.30 | 0.19 | 0.21 |
| **926** | 88 | 56 | 25.00 | 30.36 | 44.64 | 0.16 | 0.19 | 0.28 |
| **927** | 26 | 13 | 46.15 | 30.77 | 23.08 | 0.23 | 0.15 | 0.12 |
| **928** | 91 | 60 | 73.33 | 16.67 | 10.00 | 0.48 | 0.11 | 0.07 |
| **929** | 166 | 87 | 47.13 | 18.39 | 34.48 | 0.25 | 0.10 | 0.18 |
| **930** | 135 | 86 | 58.14 | 23.26 | 18.60 | 0.37 | 0.15 | 0.12 |
| **931** | 116 | 58 | 77.59 | 8.62 | 13.79 | 0.39 | 0.04 | 0.07 |
| **932** | 157 | 116 | 58.62 | 19.83 | 21.55 | 0.43 | 0.15 | 0.16 |
| **933** | 67 | 23 | 47.83 | 39.13 | 13.04 | 0.16 | 0.13 | 0.04 |
| **934** | 92 | 49 | 53.06 | 20.41 | 26.53 | 0.28 | 0.11 | 0.14 |
| **935** | 78 | 87 | 66.67 | 6.90 | 26.44 | 0.74 | 0.08 | 0.29 |
| **936** | 108 | 78 | 61.54 | 16.67 | 21.79 | 0.44 | 0.12 | 0.16 |
| **937** | 42 | 19 | 63.16 | 15.79 | 21.05 | 0.29 | 0.07 | 0.10 |
| **938** | 80 | 27 | 74.07 | 14.81 | 11.11 | 0.25 | 0.05 | 0.04 |
| **939** | 102 | 91 | 68.13 | 13.19 | 18.68 | 0.61 | 0.12 | 0.17 |

**Table 6 cont. Student Erasures by School (Reading Content Area)**

| | N | Total Erasures | % WR | % RW | % WW | Mean WR per student | Mean RW per student | Mean WW per student |
|---|---|---|---|---|---|---|---|---|
| **940** | 89 | 56 | 42.86 | 32.14 | 25.00 | 0.27 | 0.20 | 0.16 |
| **941** | 85 | 73 | 42.47 | 12.33 | 45.21 | 0.36 | 0.11 | 0.39 |
| **942** | 115 | 54 | 59.26 | 22.22 | 18.52 | 0.28 | 0.10 | 0.09 |
| **943** | 61 | 65 | 70.77 | 16.92 | 12.31 | 0.75 | 0.18 | 0.13 |
| **944** | 20 | 15 | 80.00 | 6.67 | 13.33 | 0.60 | 0.05 | 0.10 |
| **945** | 154 | 89 | 69.66 | 13.48 | 16.85 | 0.40 | 0.08 | 0.10 |
| **946** | 87 | 81 | 43.21 | 28.40 | 28.40 | 0.40 | 0.26 | 0.26 |
| **947** | 64 | 51 | 68.63 | 5.88 | 25.49 | 0.55 | 0.05 | 0.20 |
| **948** | 150 | 64 | 46.88 | 26.56 | 26.56 | 0.20 | 0.11 | 0.11 |
| **Total** | 4248 | 2772 | 56.39 | 18.98 | 24.64 | 0.37 | 0.12 | 0.16 |

***Note**. WR: wrong to right, RW: right to wrong, and WW: wrong to wrong There are 35 multiple choice items in the mathematics content area. This differs from the reading content area which contains 18 multiple choice items.

**Table 7. Facets Summary Statistics (Mathematics Content Area)**

| | Facets Summary Statistics | | | |
|---|---|---|---|---|
| | Persons | Items | PrePost | Schools |
| Mean Estimate (SD) | 1.01 (1.41) | .00 (.53) | .00 (.07) | .00 (.33) |
| Reliability of Estimates | .90 | >.99 | .99 | .99 |
| Infit MNSQ (SD) | 1.00 (.07) | 1.00 (.09) | 1.00 (.01) | 1.00 (.02) |
| Outfit MNSQ (SD) | 1.00 (.19) | 1.00 (.17) | 1.00 (.02) | 1.00 (.03) |
| Chi-Square | 40945.6* | 12404.1* | 261.7* | 5235.2* |
| Degrees of Freedom | 4246 | 34 | 1 | 47 |

*p<.01

**Table 8. Facets Summary Statistics (Reading Content Area)**

| | Facets Summary Statistics | | | |
|---|---|---|---|---|
| | Persons | Items | PrePost | Schools |
| Mean Estimate (SD) | 1.07 (1.49) | .00 (.59) | .00 (.04) | .00 (.26) |
| Reliability of Estimates | .83 | >.99 | .96 | .96 |
| | | | | |
| Infit MNSQ (SD) | 1.00 (.12) | 1.00 (.10) | 1.00 (.00) | 1.00 (.02) |
| Outfit MNSQ (SD) | .98 (.29) | .98 (.16) | .98 (.01) | .99 (.05) |
| | | | | |
| Chi-Square | 22453.2* | 7784.6* | 45.3* | 1628.6* |
| Degrees of Freedom | 4246 | 17 | 1 | 47 |

*p<.01

**Table 9. Fit Statistics for Pre and Post Erasure by School (Mathematics Content Area)**

| | Pre | | | | | | | | Post | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **School** | Measure | S.E. | InfitMS | InfitZ | OutfitMS | OutfitZ | Z | Discrim | Measure | S.E. | InfitMS | InfitZ | OutfitMS | OutfitZ | Z | Discrim |
| **901** | -0.03 | 0.05 | 1.02 | 0.82 | 1.03 | 0.69 | -0.005 | 0.96 | -0.04 | 0.05 | 1.03 | 1.16 | 1.04 | 0.83 | -0.006 | 0.94 |
| **902** | 0.08 | 0.04 | 0.99 | -0.47 | 0.98 | -0.43 | 0.001 | 1.02 | 0.13 | 0.04 | 0.98 | -0.91 | 1.00 | 0.01 | 0.001 | 1.03 |
| **903** | 0.19 | 0.11 | 0.99 | -0.22 | 0.97 | -0.19 | 0.002 | 1.02 | 0.41 | 0.12 | 1.02 | 0.29 | 1.00 | 0.02 | 0.000 | 0.98 |
| **904** | 0.21 | 0.05 | 0.99 | -0.42 | 1.00 | 0.06 | 0.001 | 1.02 | 0.28 | 0.06 | 0.99 | -0.38 | 0.98 | -0.30 | 0.003 | 1.02 |
| **905** | 0.20 | 0.05 | 1.02 | 0.73 | 1.02 | 0.35 | 0.000 | 0.98 | 0.15 | 0.05 | 1.02 | 0.73 | 1.02 | 0.39 | 0.000 | 0.97 |
| **906** | 0.08 | 0.05 | 1.03 | 1.53 | 1.04 | 0.91 | -0.007 | 0.94 | 0.01 | 0.05 | 1.03 | 1.54 | 1.05 | 1.09 | -0.006 | 0.94 |
| **907** | -0.23 | 0.04 | 1.01 | 0.75 | 1.01 | 0.44 | -0.001 | 0.97 | -0.27 | 0.04 | 1.01 | 0.54 | 1.01 | 0.19 | -0.001 | 0.98 |
| **908** | -0.59 | 0.05 | 1.00 | -0.11 | 0.99 | -0.18 | 0.001 | 1.01 | -0.47 | 0.06 | 1.02 | 0.69 | 1.02 | 0.55 | 0.000 | 0.96 |
| **909** | 0.14 | 0.05 | 1.01 | 0.42 | 1.05 | 1.26 | -0.010 | 0.97 | 0.21 | 0.05 | 1.01 | 0.28 | 1.04 | 0.77 | -0.009 | 0.98 |
| **910** | 0.31 | 0.05 | 1.01 | 0.39 | 1.03 | 0.59 | 0.004 | 0.99 | 0.32 | 0.05 | 1.01 | 0.55 | 1.03 | 0.56 | 0.002 | 0.98 |
| **911** | -0.22 | 0.04 | 0.98 | -1.26 | 0.95 | -1.74 | 0.004 | 1.07 | -0.23 | 0.04 | 0.98 | -1.19 | 0.95 | -1.50 | 0.003 | 1.06 |
| **912** | -0.37 | 0.03 | 1.00 | -0.34 | 0.99 | -0.45 | 0.001 | 1.01 | -0.41 | 0.03 | 0.99 | -0.64 | 0.99 | -0.55 | 0.000 | 1.02 |
| **913** | -0.27 | 0.06 | 0.97 | -1.11 | 0.95 | -1.40 | 0.006 | 1.07 | -0.28 | 0.06 | 0.98 | -0.98 | 0.96 | -1.07 | 0.006 | 1.06 |
| **914** | -0.69 | 0.05 | 1.02 | 0.99 | 1.01 | 0.37 | -0.001 | 0.96 | -0.82 | 0.05 | 1.01 | 0.53 | 1.00 | 0.08 | -0.001 | 0.98 |
| **915** | -0.18 | 0.06 | 0.99 | -0.43 | 0.97 | -0.78 | 0.004 | 1.03 | -0.19 | 0.06 | 0.99 | -0.28 | 0.97 | -0.73 | 0.005 | 1.02 |
| **916** | 0.55 | 0.04 | 0.99 | -0.42 | 0.96 | -0.99 | 0.009 | 1.01 | 0.51 | 0.04 | 0.99 | -0.25 | 0.96 | -0.79 | 0.012 | 1.01 |
| **917** | 0.25 | 0.05 | 0.98 | -0.67 | 0.96 | -0.79 | 0.000 | 1.03 | 0.26 | 0.05 | 0.99 | -0.56 | 0.97 | -0.62 | -0.002 | 1.02 |
| **918** | 0.33 | 0.05 | 0.98 | -0.62 | 0.98 | -0.30 | 0.009 | 1.02 | 0.34 | 0.05 | 0.98 | -0.63 | 0.99 | -0.24 | 0.009 | 1.02 |
| **919** | -0.40 | 0.04 | 1.00 | -0.07 | 0.99 | -0.26 | 0.000 | 1.00 | -0.43 | 0.04 | 1.00 | -0.08 | 1.00 | -0.11 | 0.000 | 1.00 |
| **920** | -0.03 | 0.05 | 1.01 | 0.29 | 1.03 | 0.60 | -0.002 | 0.99 | -0.06 | 0.05 | 1.01 | 0.28 | 1.02 | 0.50 | -0.002 | 0.99 |
| **921** | -0.06 | 0.04 | 0.97 | -1.60 | 0.96 | -1.52 | -0.002 | 1.07 | 0.03 | 0.04 | 0.98 | -1.32 | 0.95 | -1.43 | -0.001 | 1.05 |
| **922** | 0.43 | 0.15 | 0.97 | -0.41 | 0.91 | -0.49 | 0.009 | 1.05 | 0.31 | 0.16 | 0.98 | -0.22 | 0.91 | -0.51 | 0.020 | 1.04 |
| **923** | 0.17 | 0.04 | 1.00 | -0.22 | 1.00 | 0.05 | 0.005 | 1.01 | 0.13 | 0.04 | 0.99 | -0.29 | 0.99 | -0.20 | 0.006 | 1.01 |
| **924** | -0.39 | 0.03 | 1.00 | -0.33 | 1.00 | 0.14 | 0.003 | 1.01 | -0.45 | 0.04 | 1.00 | -0.22 | 1.00 | 0.06 | 0.005 | 1.01 |
| **925** | -0.01 | 0.04 | 1.00 | -0.14 | 0.99 | -0.25 | 0.004 | 1.01 | 0.02 | 0.04 | 1.00 | 0.16 | 1.00 | -0.03 | 0.004 | 1.00 |
| **926** | -0.39 | 0.04 | 1.01 | 0.50 | 1.01 | 0.40 | 0.003 | 0.98 | -0.40 | 0.04 | 1.00 | 0.30 | 1.01 | 0.20 | 0.003 | 0.99 |
| **927** | -0.08 | 0.08 | 1.00 | -0.08 | 0.97 | -0.49 | 0.005 | 1.02 | -0.09 | 0.08 | 0.99 | -0.31 | 0.96 | -0.64 | 0.007 | 1.03 |
| **928** | 0.47 | 0.05 | 1.01 | 0.34 | 1.03 | 0.52 | 0.004 | 0.99 | 0.48 | 0.05 | 1.01 | 0.22 | 1.03 | 0.49 | 0.008 | 0.99 |

**Table 9 cont. Fit Statistics for Pre and Post Erasure by School (Mathematics Content Area)**

| | Pre | | | | | | | | Post | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **School** | Measure | S.E. | InfitMS | InfitZ | OutfitMS | OutfitZ | Z | Discrim | Measure | S.E. | InfitMS | InfitZ | OutfitMS | OutfitZ | Z | Discrim |
| **929** | -0.18 | 0.03 | 1.01 | 0.80 | 1.02 | 0.73 | -0.001 | 0.97 | -0.21 | 0.03 | 1.01 | 0.78 | 1.01 | 0.50 | 0.000 | 0.97 |
| **930** | 0.49 | 0.04 | 0.98 | -0.95 | 0.95 | -0.98 | 0.011 | 1.03 | 0.47 | 0.04 | 0.98 | -0.74 | 0.95 | -0.95 | 0.012 | 1.03 |
| **931** | -0.02 | 0.04 | 1.00 | -0.18 | 0.99 | -0.41 | 0.002 | 1.01 | 0.03 | 0.04 | 1.00 | -0.14 | 0.98 | -0.53 | 0.004 | 1.01 |
| **932** | 0.03 | 0.03 | 1.00 | -0.10 | 1.01 | 0.27 | 0.001 | 1.00 | 0.05 | 0.03 | 1.00 | 0.07 | 1.01 | 0.43 | 0.000 | 0.99 |
| **933** | 0.12 | 0.05 | 1.03 | 1.17 | 1.04 | 0.93 | 0.000 | 0.95 | 0.11 | 0.05 | 1.03 | 1.10 | 1.03 | 0.56 | 0.001 | 0.96 |
| **934** | -0.18 | 0.04 | 0.98 | -1.16 | 1.00 | 0.02 | -0.001 | 1.03 | -0.27 | 0.04 | 0.98 | -0.96 | 1.01 | 0.33 | 0.000 | 1.02 |
| **935** | -0.30 | 0.04 | 1.00 | 0.21 | 1.00 | 0.09 | 0.007 | 0.99 | -0.31 | 0.04 | 1.00 | -0.02 | 0.99 | -0.31 | 0.007 | 1.00 |
| **936** | -0.27 | 0.04 | 1.01 | 0.46 | 1.01 | 0.35 | 0.003 | 0.98 | -0.30 | 0.04 | 1.01 | 0.38 | 1.02 | 0.68 | 0.003 | 0.99 |
| **937** | -0.09 | 0.06 | 1.01 | 0.49 | 1.06 | 1.21 | -0.005 | 0.96 | -0.05 | 0.06 | 1.01 | 0.28 | 1.01 | 0.14 | 0.001 | 0.99 |
| **938** | 0.32 | 0.05 | 1.04 | 1.59 | 1.04 | 0.81 | -0.005 | 0.95 | 0.27 | 0.05 | 1.03 | 1.26 | 1.02 | 0.37 | -0.001 | 0.96 |
| **939** | 0.13 | 0.04 | 1.02 | 1.16 | 1.00 | -0.05 | 0.006 | 0.97 | 0.19 | 0.04 | 1.02 | 0.92 | 1.02 | 0.42 | 0.004 | 0.97 |
| **940** | 0.12 | 0.04 | 1.02 | 0.84 | 1.05 | 1.20 | -0.005 | 0.96 | 0.09 | 0.05 | 1.02 | 0.81 | 1.04 | 0.99 | -0.003 | 0.97 |
| **941** | -0.73 | 0.04 | 1.01 | 0.51 | 1.01 | 0.55 | 0.005 | 0.98 | -0.66 | 0.04 | 1.00 | 0.23 | 0.99 | -0.24 | 0.004 | 0.99 |
| **942** | 0.44 | 0.04 | 0.99 | -0.38 | 1.05 | 1.12 | -0.006 | 1.00 | 0.46 | 0.04 | 1.00 | -0.17 | 1.04 | 0.74 | -0.004 | 1.00 |
| **943** | 0.26 | 0.05 | 1.02 | 0.73 | 1.03 | 0.58 | -0.004 | 0.97 | 0.25 | 0.06 | 1.01 | 0.53 | 1.05 | 0.88 | -0.004 | 0.97 |
| **944** | 0.20 | 0.09 | 1.01 | 0.30 | 1.00 | 0.00 | -0.001 | 0.99 | 0.29 | 0.10 | 1.01 | 0.31 | 1.01 | 0.11 | -0.002 | 0.99 |
| **945** | -0.24 | 0.03 | 0.99 | -1.02 | 1.00 | 0.14 | -0.003 | 1.03 | -0.23 | 0.03 | 0.99 | -0.94 | 1.00 | -0.04 | -0.002 | 1.03 |
| **946** | -0.38 | 0.04 | 0.99 | -0.69 | 0.99 | -0.32 | 0.001 | 1.02 | -0.32 | 0.04 | 0.98 | -0.84 | 0.99 | -0.23 | 0.002 | 1.03 |
| **947** | 0.19 | 0.05 | 0.97 | -1.13 | 0.93 | -1.29 | 0.008 | 1.05 | 0.12 | 0.06 | 0.98 | -0.83 | 0.95 | -0.92 | 0.009 | 1.04 |
| **948** | 0.58 | 0.04 | 1.00 | -0.20 | 0.96 | -0.81 | 0.006 | 1.01 | 0.54 | 0.04 | 1.00 | -0.02 | 0.98 | -0.32 | 0.005 | 1.00 |

**Table 10. Fit Statistics for Pre and Post Erasure by School (Reading Content Area)**

| | Pre | | | | | | | | Post | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **School** | Measure | S.E. | InfitMS | InfitZ | OutfitMS | OutfitZ | Discrim | Z | Measure | S.E. | InfitMS | InfitZ | OutfitMS | OutfitZ | Discrim | Z |
| **901** | 0.03 | 0.08 | 1.01 | 0.28 | 0.99 | -0.13 | 0.99 | 0.006 | 0.01 | 0.08 | 1.01 | 0.35 | 0.99 | -0.14 | 0.98 | 0.005 |
| **902** | 0.11 | 0.06 | 1.01 | 0.18 | 1.01 | 0.16 | 1.00 | 0.007 | 0.18 | 0.07 | 1.02 | 0.75 | 0.99 | -0.09 | 0.97 | 0.015 |
| **903** | -0.09 | 0.15 | 0.99 | -0.12 | 0.94 | -0.50 | 1.03 | 0.017 | -0.04 | 0.15 | 1.02 | 0.24 | 0.97 | -0.18 | 0.98 | 0.020 |
| **904** | 0.22 | 0.08 | 1.00 | 0.14 | 1.03 | 0.41 | 0.99 | 0.004 | 0.16 | 0.08 | 1.01 | 0.38 | 1.03 | 0.40 | 0.97 | 0.007 |
| **905** | 0.10 | 0.06 | 1.00 | -0.08 | 0.97 | -0.44 | 1.01 | 0.003 | 0.07 | 0.06 | 1.01 | 0.24 | 0.95 | -0.86 | 1.00 | 0.004 |
| **906** | -0.12 | 0.06 | 0.97 | -1.16 | 0.93 | -1.46 | 1.07 | 0.005 | -0.14 | 0.06 | 0.99 | -0.49 | 0.95 | -1.09 | 1.04 | 0.004 |
| **907** | -0.42 | 0.05 | 1.01 | 0.34 | 1.02 | 0.53 | 0.98 | 0.010 | -0.43 | 0.06 | 1.00 | 0.06 | 1.00 | 0.12 | 1.00 | 0.012 |
| **908** | -0.58 | 0.08 | 1.02 | 0.67 | 1.05 | 1.12 | 0.93 | 0.004 | -0.50 | 0.08 | 1.03 | 1.00 | 1.05 | 0.92 | 0.92 | 0.005 |
| **909** | 0.42 | 0.08 | 0.98 | -0.57 | 0.93 | -0.99 | 1.04 | 0.011 | 0.41 | 0.08 | 0.98 | -0.49 | 0.91 | -1.14 | 1.04 | 0.015 |
| **910** | 0.10 | 0.07 | 1.00 | -0.12 | 0.97 | -0.57 | 1.02 | 0.017 | 0.08 | 0.07 | 0.99 | -0.36 | 0.96 | -0.71 | 1.03 | 0.021 |
| **911** | -0.20 | 0.06 | 1.02 | 0.82 | 1.01 | 0.34 | 0.96 | -0.001 | -0.18 | 0.06 | 1.02 | 0.62 | 1.03 | 0.61 | 0.96 | -0.003 |
| **912** | -0.04 | 0.05 | 1.00 | 0.07 | 0.97 | -0.80 | 1.01 | 0.010 | -0.05 | 0.05 | 1.01 | 0.30 | 0.96 | -0.92 | 1.00 | 0.012 |
| **913** | -0.14 | 0.09 | 1.02 | 0.42 | 1.05 | 0.69 | 0.96 | 0.010 | -0.17 | 0.09 | 1.02 | 0.58 | 1.05 | 0.66 | 0.96 | 0.015 |
| **914** | -0.51 | 0.07 | 1.02 | 0.51 | 1.05 | 0.90 | 0.96 | 0.015 | -0.50 | 0.07 | 1.02 | 0.57 | 1.03 | 0.56 | 0.96 | 0.013 |
| **915** | -0.13 | 0.09 | 0.99 | -0.32 | 0.97 | -0.50 | 1.04 | 0.011 | -0.13 | 0.09 | 0.99 | -0.20 | 0.97 | -0.36 | 1.03 | 0.010 |
| **916** | 0.28 | 0.05 | 0.98 | -0.91 | 0.92 | -1.64 | 1.04 | 0.021 | 0.25 | 0.05 | 0.97 | -1.23 | 0.92 | -1.63 | 1.05 | 0.022 |
| **917** | 0.17 | 0.07 | 1.01 | 0.19 | 0.99 | -0.22 | 1.00 | 0.009 | 0.26 | 0.07 | 0.97 | -0.78 | 0.89 | -1.63 | 1.06 | 0.018 |
| **918** | 0.12 | 0.06 | 1.01 | 0.29 | 0.99 | -0.26 | 0.99 | 0.008 | 0.09 | 0.06 | 1.02 | 0.64 | 1.00 | -0.04 | 0.98 | 0.009 |
| **919** | -0.30 | 0.05 | 1.01 | 0.54 | 1.00 | 0.02 | 0.98 | 0.013 | -0.32 | 0.05 | 1.01 | 0.60 | 1.00 | -0.02 | 0.98 | 0.013 |
| **920** | 0.17 | 0.08 | 0.98 | -0.63 | 0.93 | -0.94 | 1.04 | 0.010 | 0.14 | 0.08 | 0.97 | -0.69 | 0.92 | -1.07 | 1.05 | 0.011 |
| **921** | 0.16 | 0.06 | 0.97 | -1.04 | 0.93 | -1.50 | 1.06 | -0.003 | 0.15 | 0.06 | 0.97 | -0.95 | 0.92 | -1.51 | 1.06 | 0.001 |
| **922** | 0.46 | 0.22 | 0.99 | 0.00 | 1.15 | 0.74 | 0.98 | -0.020 | 0.49 | 0.23 | 0.99 | -0.03 | 1.14 | 0.65 | 0.98 | -0.020 |
| **923** | 0.09 | 0.06 | 0.99 | -0.38 | 0.97 | -0.58 | 1.02 | 0.012 | 0.08 | 0.06 | 0.99 | -0.52 | 0.96 | -0.69 | 1.03 | 0.011 |
| **924** | -0.41 | 0.05 | 1.03 | 1.74 | 1.04 | 1.22 | 0.91 | 0.016 | -0.42 | 0.05 | 1.03 | 1.71 | 1.03 | 0.75 | 0.92 | 0.014 |
| **925** | -0.10 | 0.06 | 1.04 | 1.34 | 0.99 | -0.15 | 0.95 | 0.015 | -0.12 | 0.06 | 1.03 | 1.31 | 1.00 | 0.02 | 0.95 | 0.014 |
| **926** | -0.43 | 0.06 | 1.01 | 0.57 | 1.01 | 0.40 | 0.97 | 0.012 | -0.48 | 0.06 | 1.02 | 0.92 | 1.03 | 0.83 | 0.94 | 0.013 |
| **927** | 0.03 | 0.11 | 0.97 | -0.63 | 1.00 | 0.07 | 1.05 | 0.016 | 0.11 | 0.12 | 0.97 | -0.57 | 1.02 | 0.24 | 1.04 | 0.008 |
| **928** | 0.26 | 0.07 | 1.02 | 0.58 | 1.03 | 0.43 | 0.97 | 0.010 | 0.19 | 0.07 | 1.02 | 0.51 | 1.02 | 0.28 | 0.98 | 0.017 |

**Table 10 cont. Fit Statistics for Pre and Post Erasure by School (Reading Content Area)**

| | Pre | | | | | | | | Post | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **School** | Measure | S.E. | InfitMS | InfitZ | OutfitMS | OutfitZ | Discrim | Z | Measure | S.E. | InfitMS | InfitZ | OutfitMS | OutfitZ | Discrim | Z |
| **929** | -0.09 | 0.04 | 0.99 | -0.38 | 0.98 | -0.59 | 1.01 | 0.011 | -0.10 | 0.05 | 0.99 | -0.31 | 0.97 | -0.68 | 1.01 | 0.011 |
| **930** | 0.27 | 0.06 | 1.00 | -0.09 | 0.95 | -0.95 | 1.01 | 0.017 | 0.27 | 0.06 | 1.01 | 0.22 | 1.00 | -0.06 | 1.00 | 0.015 |
| **931** | 0.15 | 0.06 | 0.97 | -1.20 | 0.91 | -1.81 | 1.06 | 0.011 | 0.15 | 0.06 | 0.98 | -0.81 | 0.91 | -1.87 | 1.05 | 0.017 |
| **932** | 0.21 | 0.05 | 1.01 | 0.49 | 1.03 | 0.67 | 0.99 | 0.004 | 0.13 | 0.05 | 1.01 | 0.44 | 1.00 | -0.01 | 0.99 | 0.013 |
| **933** | 0.18 | 0.08 | 0.99 | -0.35 | 0.98 | -0.29 | 1.03 | 0.020 | 0.12 | 0.08 | 0.99 | -0.31 | 0.99 | -0.04 | 1.02 | 0.018 |
| **934** | -0.26 | 0.06 | 1.02 | 0.63 | 0.99 | -0.18 | 0.97 | 0.013 | -0.30 | 0.06 | 1.01 | 0.53 | 0.99 | -0.23 | 0.98 | 0.014 |
| **935** | -0.22 | 0.06 | 1.02 | 0.83 | 1.03 | 0.71 | 0.96 | 0.002 | -0.13 | 0.06 | 1.03 | 1.11 | 1.04 | 0.75 | 0.94 | 0.002 |
| **936** | -0.15 | 0.05 | 1.05 | 1.83 | 1.07 | 1.52 | 0.91 | 0.007 | -0.16 | 0.06 | 1.05 | 1.81 | 1.06 | 1.24 | 0.91 | 0.009 |
| **937** | 0.20 | 0.10 | 1.00 | 0.00 | 0.92 | -0.83 | 1.02 | 0.019 | 0.23 | 0.10 | 1.00 | -0.04 | 0.91 | -0.93 | 1.02 | 0.021 |
| **938** | 0.30 | 0.07 | 1.02 | 0.57 | 1.01 | 0.23 | 0.98 | 0.011 | 0.23 | 0.07 | 1.02 | 0.59 | 1.02 | 0.29 | 0.98 | 0.014 |
| **939** | 0.05 | 0.06 | 0.99 | -0.37 | 0.93 | -1.22 | 1.03 | 0.014 | 0.06 | 0.06 | 0.98 | -0.84 | 0.92 | -1.47 | 1.05 | 0.016 |
| **940** | 0.25 | 0.06 | 0.99 | -0.39 | 0.95 | -0.87 | 1.02 | 0.011 | 0.24 | 0.07 | 0.99 | -0.15 | 0.97 | -0.47 | 1.01 | 0.009 |
| **941** | -0.61 | 0.06 | 1.03 | 0.98 | 1.05 | 1.26 | 0.94 | 0.015 | -0.60 | 0.06 | 1.03 | 1.06 | 1.06 | 1.37 | 0.93 | 0.015 |
| **942** | 0.21 | 0.06 | 0.97 | -1.08 | 0.94 | -1.17 | 1.05 | 0.017 | 0.18 | 0.06 | 0.97 | -1.02 | 0.93 | -1.22 | 1.05 | 0.018 |
| **943** | 0.23 | 0.08 | 1.00 | -0.04 | 1.03 | 0.47 | 1.00 | 0.010 | 0.32 | 0.09 | 1.00 | -0.07 | 0.99 | -0.03 | 1.01 | 0.015 |
| **944** | 0.09 | 0.13 | 1.02 | 0.30 | 1.01 | 0.11 | 0.97 | 0.007 | 0.13 | 0.13 | 1.00 | 0.07 | 0.99 | -0.01 | 1.00 | 0.007 |
| **945** | 0.04 | 0.05 | 1.00 | -0.03 | 0.97 | -0.84 | 1.01 | 0.003 | 0.07 | 0.05 | 1.00 | -0.17 | 0.96 | -1.05 | 1.02 | 0.004 |
| **946** | -0.28 | 0.06 | 1.03 | 1.09 | 1.07 | 1.49 | 0.93 | 0.001 | -0.30 | 0.06 | 1.03 | 1.03 | 1.10 | 1.77 | 0.93 | -0.001 |
| **947** | 0.02 | 0.08 | 1.00 | 0.06 | 0.96 | -0.58 | 1.01 | 0.017 | 0.10 | 0.08 | 1.00 | 0.00 | 0.95 | -0.63 | 1.01 | 0.015 |
| **948** | 0.20 | 0.05 | 1.00 | -0.16 | 0.97 | -0.54 | 1.01 | 0.014 | 0.18 | 0.05 | 0.99 | -0.28 | 0.95 | -0.95 | 1.02 | 0.017 |

**Table 11. Mapping of Research Questions and Analysis**

| Research Question | Analysis |
|---|---|
| 1. What amount of variance in aberrant responds is accounted for at the school level? | Investigation of variance component |
| 2. Do select student- and school-level factors predict aberrant responding? | Investigation of the regression coefficients |

**Table 12. Distribution of Aberrance**

| | **Math** | | | | **Reading** | | | |
|---|---|---|---|---|---|---|---|---|
| | Outfit | | Infit | | Outfit | | Infit | |
| | Non-Aberrant | Aberrant | Non-Aberrant | Aberrant | Non-Aberrant | Aberrant | Non-Aberrant | Aberrant |
| Frequency (N) | 3802 | 446 | 4215 | 33 | 3524 | 724 | 3984 | 264 |
| Frequency (%) | 89.5 | 10.5 | 99.2 | .8 | 83 | 17 | 93.8 | 6.2 |

**Note: Non-aberrant: Outift/Infit < 1.20, Aberrant: Outfit/Infit ≥ 1.20*

**Table 13. Explanatory variables defined**

| Variable | Student-Level | School-Level |
|---|---|---|
| DEMOGRAPHIC | | |
| Gender | 0=Male, 1=Female | Percentage of males in each school |
| Economic Status | 0=No, 1=Yes | Percentage of students that are economically disadvantaged in each school |
| BEHAVIORAL | | |
| Erasure Behavior | Total Erasures per student | Mean number of erasures for school |
| Mathematics/ Reading Proficiency | 1=Advanced Proficient, 2=Proficient, 3=Partially Proficient | School level proficiency |

**Table 14. Summary Statistics from Facets Analyses**

| | **Mathematics** | | **Reading** | |
|---|---|---|---|---|
| | **Students** (N=4248) | **Items** (n=35) | **Students** (N=4248) | **Items** (n=18) |
| Mean Estimate (SD) | 1.01 (1.49) | .00 (.53) | 1.11 (1.47) | .00 (.59) |
| Reliability of Estimates | .84 | .99 | 0.72 | >.99 |
| Infit MNSQ (SD) | 1.00 (.07) | 1.00 (.10) | 1.00 (.12) | 1.00 (.10) |
| Outfit MNSQ (SD) | 1.00 (.20) | 1.00 (.18) | .98 (.29) | .98 (.17) |
| Chi-Square | 23672.9* | 6048.4* | 12705.3* | 3844.3* |
| Degrees of Freedom | 4247 | 34 | 4246 | 17 |

*p<.01

**Table 15. Means and Standard Deviations**

| | **Student** | | **School** | |
|---|---|---|---|---|
| | M | SD | M | SD |
| **Demographic** | | | | |
| Female | 0.480 | 0.500 | 0.485 | 0.063 |
| Economically Disadvantaged | 0.290 | 0.453 | 0.294 | 0.269 |
| **Behavior** | | | | |
| Mathematics Total Erasures | 1.580 | 1.767 | 1.681 | 0.646 |
| Reading Total Erasures | 0.653 | 1.087 | 0.665 | 0.275 |
| Mathematics Proficiency | | | | |
| Advanced Proficient | 0.399 | 0.490 | 0.402 | 0.180 |
| Proficient | 0.406 | 0.491 | 0.410 | 0.106 |
| Partially Proficient | 0.182 | 0.386 | 0.176 | 0.127 |
| Reading Proficiency | | | | |
| Advanced Proficient | 0.073 | 0.260 | 0.071 | 0.074 |
| Proficient | 0.560 | 0.496 | 0.560 | 0.125 |
| Partially Proficient | 0.353 | 0.478 | 0.357 | 0.167 |
| **Dependent Variables** | | | | |
| $\text{Outfit}_{\text{Math}}$ Count | 0.105 | 0.310 | 0.106 | 0.047 |
| $\text{Infit}_{\text{Math}}$ Count | 0.008 | 0.088 | 0.008 | 0.010 |
| $\text{Outfit}_{\text{Read}}$ Count | 0.170 | 0.376 | 0.179 | 0.072 |
| $\text{Infit}_{\text{Read}}$ Count | 0.062 | 0.241 | 0.061 | 0.036 |

**Table 16. Percentage of misfitting persons with Outfit MSE or Infit MSE above 1.20**

| | Mathematics | | Reading | |
|---|---|---|---|---|
| **Variable** | **Outfit MSE** %(SD) | **Infit MSE** %(SD) | **Outfit MSE** %(SD) | **Infit MSE** %(SD) |
| **Gender** | | | | |
| Male | 11.4 (0.318) | 1.1 (0.102) | 17.9 (0.383) | 0.066 (0.248) |
| Female | 9.6 (0.294) | 0.5 (0.070) | 16.2 (0.369) | 0.059 (0.235) |
| **Economic Status** | | | | |
| No | 11.1 (0.314) | 0.6 (0.80) | 15.4 (0.361) | 4.6 (0.210) |
| Yes | 9.1 (0.288) | 1.1 (0.106) | 21.1 (0.408) | 10.2 (0.303) |
| **Proficiency Level** | | | | |
| Advanced Proficient | 17.4 (0.380) | 0.12 (0.034) | 10.0 (0.301) | 0.32 (0.057) |
| Proficient | 5.2 (0.221) | 0.9 (0.093) | 14.9 (0.356) | 3.03 (0.171) |
| Partially Proficient | 7.8 (0.268) | 1.9 (0.138) | 22.5 (0.418) | 12.6 (0.332) |
| **Total** | **10.5%** | **0.8 (0.088)** | **17.0 (0.376)** | **6.2 (0.241)** |

**Table 17. Parameter Estimates for the Outfit Two-Level Model for Mathematics Content Area**

| | **$Outfit_A$** | | | **$Outfit_B$** | | | **$Outfit_C$** | | |
|---|---|---|---|---|---|---|---|---|---|
| | β | se | Sig. | β | se | Sig. | β | se | Sig. |
| Intercept | **-2.132** | **0.061** | **<0.001** | **-2.139** | **0.059** | **<0.001** | **-2.286** | **0.060** | **<0.001** |
| Gender | | | | **-0.209** | **0.101** | **0.039** | -0.186 | 0.103 | 0.072 |
| Eco. Status | | | | -0.198 | 0.120 | 0.099 | 0.132 | 0.128 | 0.301 |
| Erasures | | | | | | | 0.026 | 0.031 | 0.399 |
| Mathematics Proficiency | | | | | | | | | |
| Advanced | | | | | | | **1.397** | **0.131** | **<0.001** |
| Proficient | | | | | | | | | |
| Partially Proficient | | | | | | | **0.417** | **0.175** | **0.017** |
| Variance Components | **0.051** | | **0.020** | 0.038 | | 0.059 | 0.011 | | 0.296 |

*Note: Bold notates statistically significant at .05 level

**Table 18. Parameter Estimates for the Infit Two-Level Model for Mathematics Content Area**

| | **Infit$_A$** | | | **Infit$_B$** | | | **Infit$_C$** | | |
|---|---|---|---|---|---|---|---|---|---|
| | β | se | Sig. | β | se | Sig. | β | se | Sig. |
| Intercept | **-4.866** | **0.177** | **<0.001** | **-5.010** | **0.203** | **<0.001** | **-5.578** | **0.326** | **<0.001** |
| Gender | | | | **-0.902** | **0.395** | **0.022** | **-0.966** | **-2.438** | **0.015** |
| Eco. Status | | | | 0.671 | 0.359 | 0.062 | 0.021 | 0.373 | 0.956 |
| Erasures | | | | | | | **0.197** | **0.074** | **0.005** |
| Mathematics Proficiency | | | | | | | | | |
| Advanced | | | | | | | **-1.856** | **0.761** | **0.015** |
| Proficient | | | | | | | | | |
| Partially Proficient | | | | | | | **0.820** | **0.375** | **0.029** |
| Variance Components | 0.090 | | >0.500 | 0.005 | | 0.388 | 0.001 | | >0.500 |

*Note: Bold notates statistically significant at .05 level

**Table 19. Parameter Estimates for the Outfit Two-Level Model for Reading Content Area**

| | **Outfit$_A$** | | | **Outfit$_B$** | | | **Outfit$_C$** | | |
|---|---|---|---|---|---|---|---|---|---|
| | β | se | Sig. | β | se | Sig. | β | se | Sig. |
| Intercept | **-1.563** | **0.061** | **<0.001** | **-1.574** | **0.057** | **<0.001** | **-1.598** | **0.057** | **<0.001** |
| Gender | | | | -0.142 | 0.083 | 0.087 | -0.071 | 0.084 | 0.401 |
| Eco. Status | | | | **0.355** | **0.095** | **<0.001** | **0.205** | **0.098** | **0.036** |
| Erasures | | | | | | | **0.092** | **0.035** | **0.008** |
| Reading Proficiency | | | | | | | | | |
| Advanced | | | | | | | **-0.432** | **0.201** | **0.031** |
| Proficient | | | | | | | | | |
| Partially Proficient | | | | | | | **0.404** | **0.091** | **<0.001** |
| | | | | | | | | | |
| Variance Components | **0.088** | | **<0.001** | **0.067** | | **<0.001** | **0.062** | | **<0.001** |

*Note: Bold notates statistically significant at .05 level

**Table 20. Parameter Estimates for the Infit Two-Level Model for Reading Content Area**

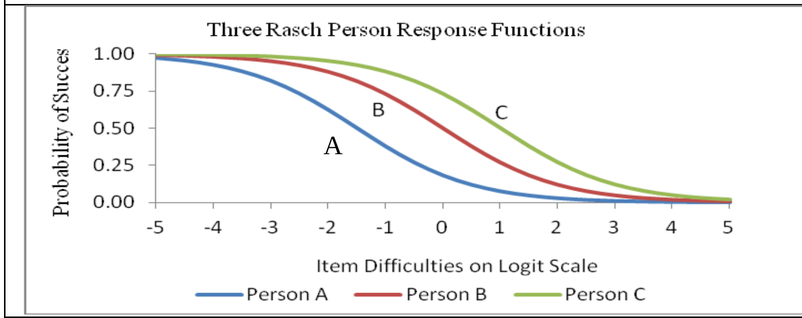
| | **$Infit_A$** | | | **$Infit_B$** | | | **$Infit_C$** | | |
|---|---|---|---|---|---|---|---|---|---|
| | β | Se | Sig. | β | se | Sig. | β | se | Sig. |
| Intercept | **-2.726** | **0.085** | **<0.001** | **-2.782** | **0.072** | **<0.001** | **-3.106** | **0.105** | **<0.001** |
| Gender | | | | -0.154 | 0.129 | 0.232 | 0.057 | 0.132 | 0.664 |
| Eco. Status | | | | **0.858** | **0.131** | **<0.001** | **0.394** | **0.136** | **0.004** |
| Erasures | | | | | | | **0.107** | **0.050** | **0.031** |
| Reading Proficiency | | | | | | | | | |
| Advanced | | | | | | | **-2.193** | **1.009** | **0.030** |
| Proficient | | | | | | | | | |
| Partially Proficient | | | | | | | **1.405** | **0.149** | **<0.001** |
| Variance Components | **0.130** | | **0.004** | 0.019 | | 0.468 | 0.001 | | >0.500 |

*Note: Bold notates statistically significant at .05 level

**Figure 1. Impact of Crossing Person Response Functions**

## *Item Invariant Measurement*

### Panel A



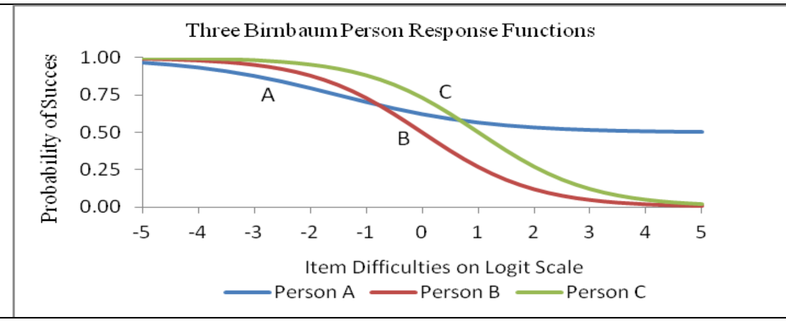

### Panel C

Three persons with same order on latent variable

| Achievement | | | |
|---|---|---|---|
| High | C | C | C |
| Medium | B | B | B |
| Low | A | A | A |
| | -2 | 0 | +2 |
| Item subsets: | Easy | Medium | Hard |

## *Item Variant Measurement*

### Panel B




### Panel D

Three persons with different orders on latent variable

| Achievement | | | |
|---|---|---|---|
| High | C | C | A |
| Medium | B | A | C |
| Low | A | B | B |
| | -2 | 0 | +2 |
| Item subsets: | Easy | Medium | Hard |

**Figure 2. Illustration of the Creation of Pre-Erasure Strings**

| | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Item 8 | Item 9 | Item 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Erasure Behavior** | None | None | None | WR | None | None | WR | None | None | None |
| **Response Pattern (Post Erasure String)** | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 |
| **Pre Erasure String** | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |

**Figure 3. Conceptual Models**

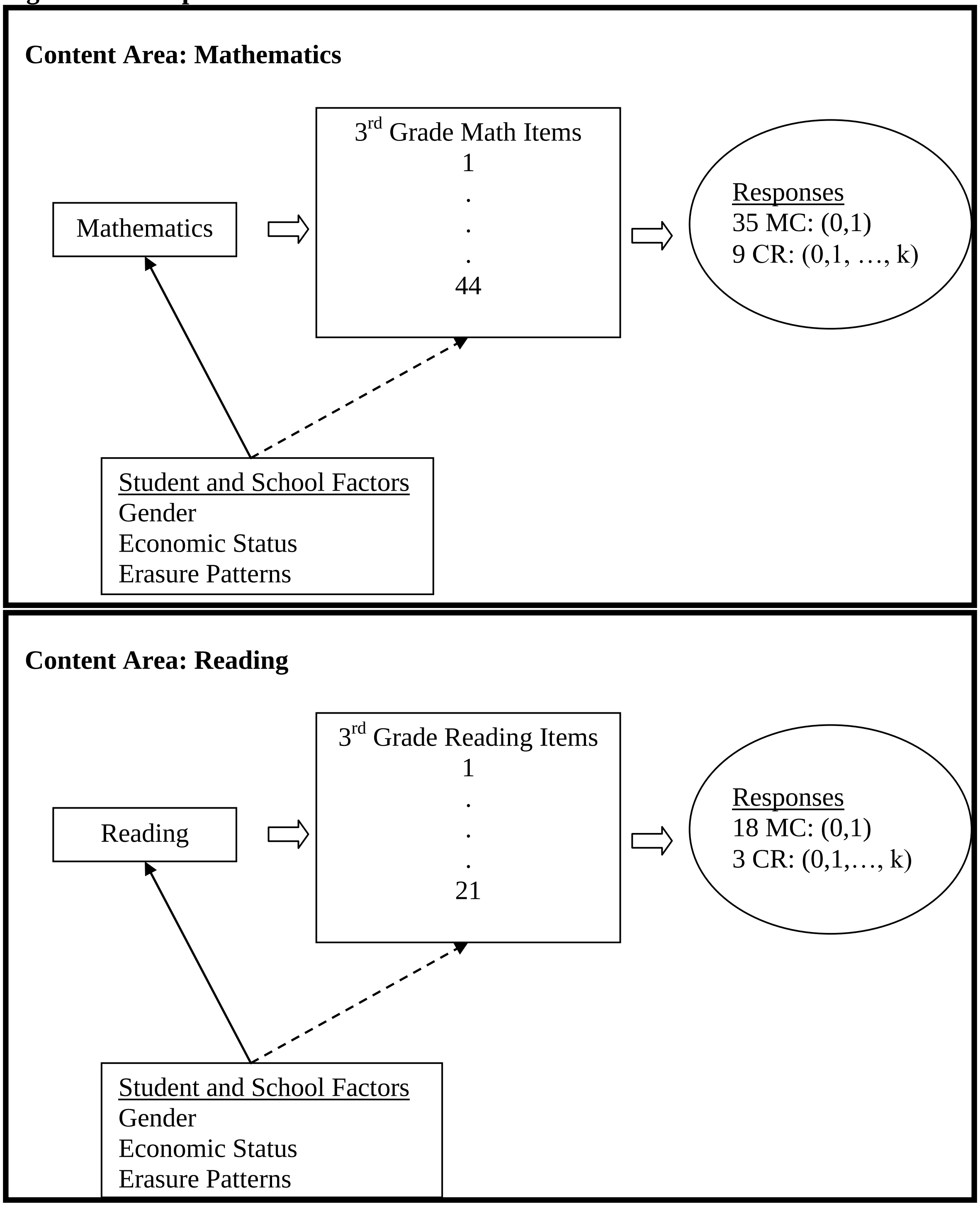


**Figure 4. Variable Map for Students and Items (Mathematics Content Area)**

```
+--------------------------------------+
|Measr|+Student     |-Item             |
|-----+------------+-------------------|
|   5 + **.         +                  |
|     | .           |                  |
|     | .           |                  |
|     | .           |                  |
|     | .           |                  |
|   4 + .           +                  |
|     | .           |                  |
|     | .           |                  |
|     | *.          |                  |
|     | *.          |                  |
|   3 + **.         +                  |
|     | ***.        |                  |
|     | ***.        |                  |
|     | *****.      |                  |
|     | ****.       |                  |
|   2 + *******.    +                  |
|     | *******.    |                  |
|     | *******.    |                  |
|     | *********.  |                  |
|     | *********.  |                  |
|   1 + *********.  + 23               |
|     | *********.  | 2  10            |
|     | *********.  | 17 19 29 32      |
|     | *********.  | 5  31 35         |
|     | ********.   | 3  4  14 25 26 27 |
*   0 * ******.     * 8  22 28 34      *
|     | ******.     | 12 13 18 20 21 24 |
|     | *****.      | 1  11 15         |
|     | ***.        | 16               |
|     | ***         | 7  9  33         |
|  -1 + *.          + 30               |
|     | *.          | 6                |
|     | .           |                  |
|     | .           |                  |
|     | .           |                  |
|  -2 + .           +                  |
|     | .           |                  |
|     | .           |                  |
|     |             |                  |
|     |             |                  |
|  -3 + .           +                  |
|     |             |                  |
|     |             |                  |
|     |             |                  |
|     |             |                  |
|  -4 +             +                  |
|     |             |                  |
|     |             |                  |
|     |             |                  |
|     |             |                  |
|  -5 + *.          +                  |
|-----+------------+-------------------|
|Measr| * = 29      |-Item             |
+--------------------------------------+
```

**Figure 5. Variable Map for School, Pre/Post Indicator, Item (Mathematics Content Area)**

```
+------------------------------------------------------------+
|Measr|+School                  |+PrePost|-Item              |
|-----+-------------------------+--------+-------------------|
|   2 +                         +        +                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        | 23                |
|   1 +                         +        +                   |
|     |                         |        |                   |
|     |                         |        | 10                |
|     |                         |        | 2  29             |
|     | 916 948                 |        | 32                |
|     | 922 928 930             |        | 17 19 35          |
|     | 942                     |        | 31                |
|     | 903 910 918 938 943     |        | 3  5              |
|     | 904 905 909 917 944 947 |        | 4  25 26 27       |
|     | 902 906 923 933 939 940 | Post   | 14 34             |
*   0 * 901 921 925 931 932     *        * 8  28             *
|     | 920 927 937             | Pre    | 22                |
|     | 911 915 929 934 945     |        | 12 13 18 20 21 24 |
|     | 907 913 935 936         |        | 1  15             |
|     | 912 919 924 926 946     |        | 11                |
|     | 908                     |        | 16                |
|     |                         |        |                   |
|     | 941                     |        | 33                |
|     | 914                     |        | 7                 |
|     |                         |        | 9                 |
|  -1 +                         +        + 30                |
|     |                         |        |                   |
|     |                         |        | 6                 |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|     |                         |        |                   |
|  -2 +                         +        +                   |
|-----+-------------------------+--------+-------------------|
|Measr|+School                  |+PrePost|-Item              |
+------------------------------------------------------------+
```

**Figure 6. Variable Map for Students and Items (Reading Content Area)**

```
+----------------------------+
|Measr|+Student     |-Item    |
|-----+------------+----------|
|   5 + *****.     +          |
|     |            |          |
|     |            |          |
|     |            |          |
|     |            |          |
|   4 + .          +          |
|     | .          |          |
|     | .          |          |
|     | .          |          |
|     | *.         |          |
|   3 + **.        +          |
|     | ****.      |          |
|     | ***.       |          |
|     | ***.       |          |
|     | ****.      |          |
|   2 + ********.  +          |
|     | *****.     |          |
|     | ********.  |          |
|     | ********.  |          |
|     | *********. | 6        |
|   1 + ********.  +          |
|     | *********. |          |
|     | *********. | 8  13 17 |
|     | ********.  | 3  10    |
|     | ******.    | 9  12 18 |
*   0 * ******.    * 1        *
|     | ****.      | 7  11 16 |
|     | ****.      |          |
|     | ***.       | 14 15    |
|     | **.        | 4  5     |
|  -1 + **.        + 2        |
|     | *.         |          |
|     | *.         |          |
|     | .          |          |
|     | .          |          |
|  -2 + .          +          |
|     | .          |          |
|     | .          |          |
|     | .          |          |
|     |            |          |
|  -3 +            +          |
|     |            |          |
|     |            |          |
|     |            |          |
|     |            |          |
|  -4 +            +          |
|     |            |          |
|     |            |          |
|     |            |          |
|     |            |          |
|  -5 + **         +          |
|-----+------------+----------|
|Measr| * = 30     |-Item     |
+----------------------------+
```

**Figure 7. Variable Map for School, Pre/Post Indicator, Items (Reading Content Area)**

```
+---------------------------------------------------------------+
|Measr|+School                              |+PrePost|-Item     |
|-----+-------------------------------------+--------+----------|
|   2 +                                     +        +          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        | 6        |
|     |                                     |        |          |
|     |                                     |        |          |
|   1 +                                     +        +          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        | 8  17    |
|     |                                     |        |          |
|     |                                     |        | 13       |
|     | 922                                 |        |          |
|     | 909                                 |        | 10       |
|     | 930 938                             |        | 3        |
|     | 916 928 943                         |        | 9        |
|     | 902 904 917 921 932 937 940 942 948 |        | 12       |
|     | 918 920 927 931 933                 |        | 18       |
|     | 905 910 923 939 944 945 947         | Post   | 1        |
*   0 * 901                                 *        *          *
|     | 903 912 929                         | Pre    |          |
|     | 906 913 915 925 936                 |        |          |
|     | 911 935                             |        | 7  11 16 |
|     | 919 934                             |        |          |
|     | 946                                 |        |          |
|     |                                     |        |          |
|     | 907 924 926                         |        |          |
|     | 908 914                             |        |          |
|     | 941                                 |        | 14       |
|     |                                     |        | 15       |
|     |                                     |        |          |
|     |                                     |        | 5        |
|     |                                     |        | 4        |
|     |                                     |        | 2        |
|  -1 +                                     +        +          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|     |                                     |        |          |
|  -2 +                                     +        +          |
|-----+-------------------------------------+--------+----------|
|Measr|+School                              |+PrePost|-Item     |
+---------------------------------------------------------------+
```

**Figure 8. Mean Wrong to Right by Total Erasures at the School Level (Mathematics Content Area)**

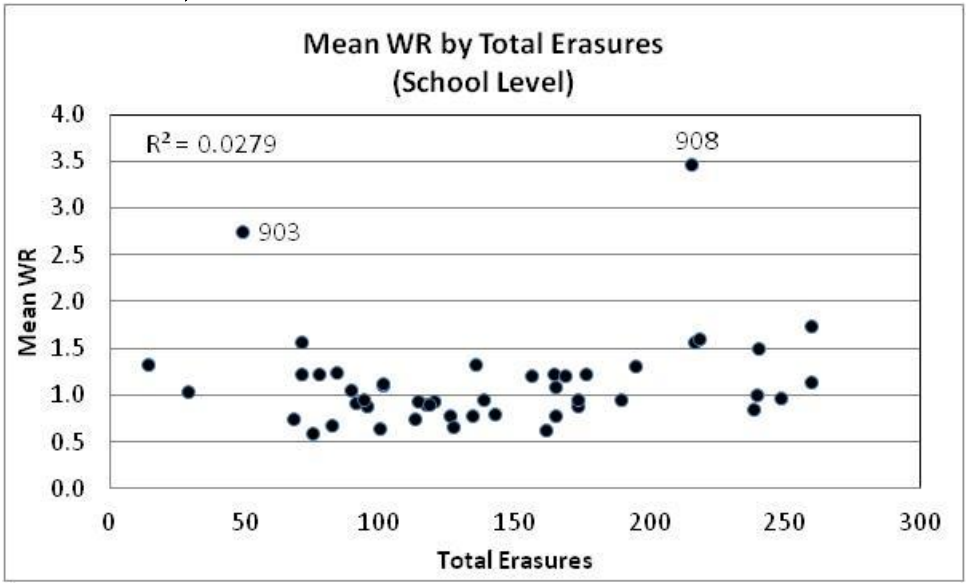


**Figure 9. Mean Post Erasure School Mathematics Achievement by Mean Pre Erasure School Mathematics Achievement**

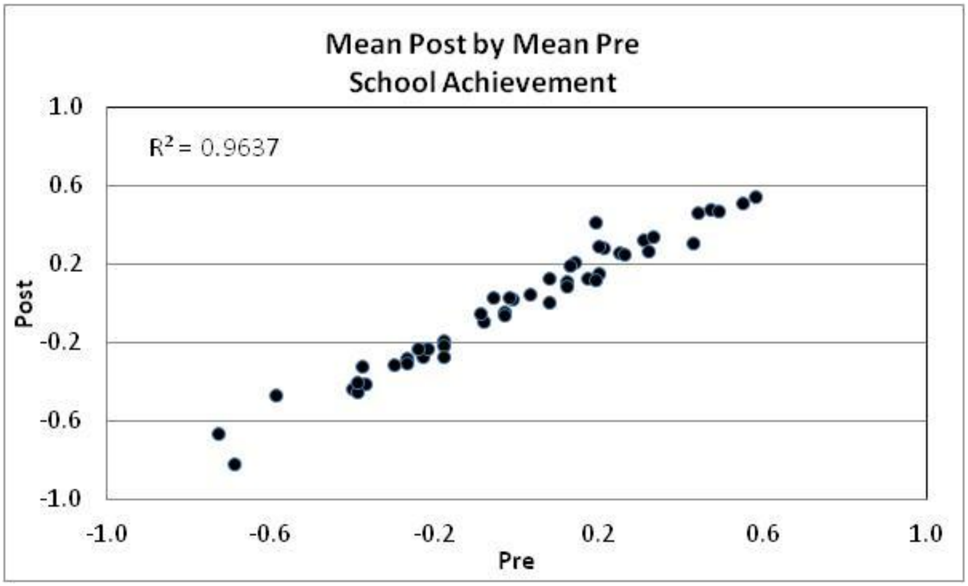


**Figure 10. Pre/Post Erasure Indicator by School Mathematics Achievement**

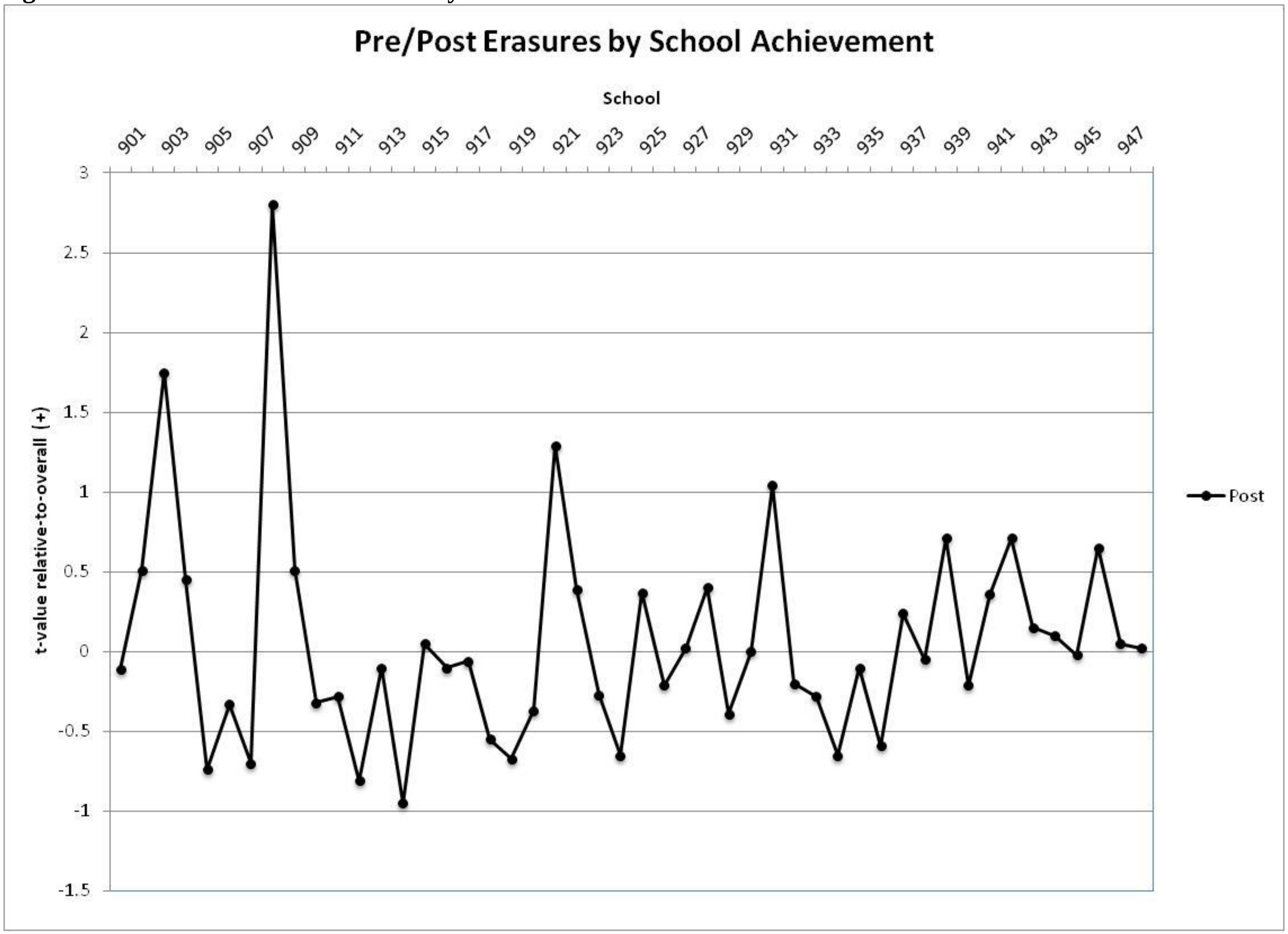


**Figure 11. Pre and Post Erasure Outfit Z by Mean Wrong to Right at the School Level (Mathematics Content Area)**

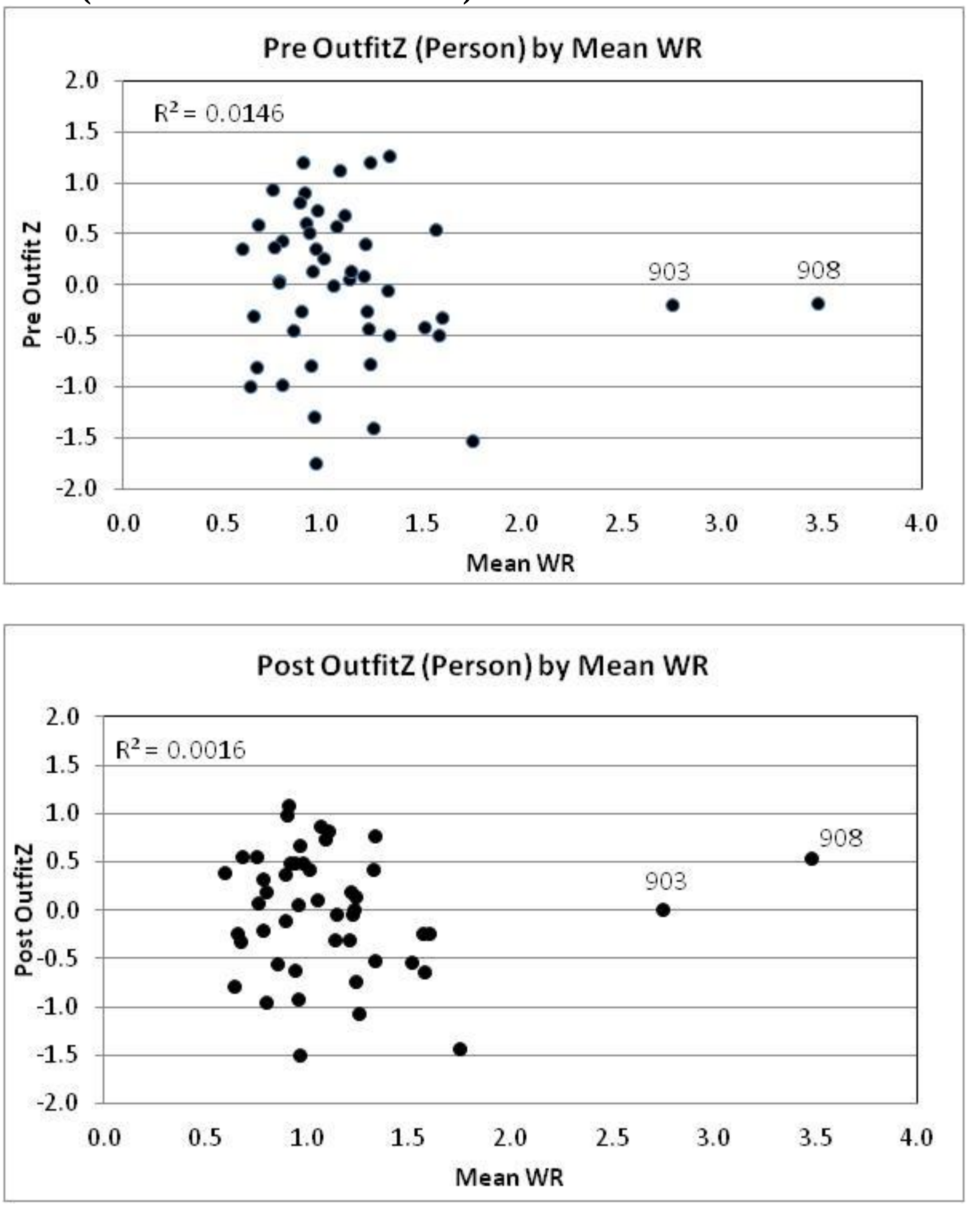
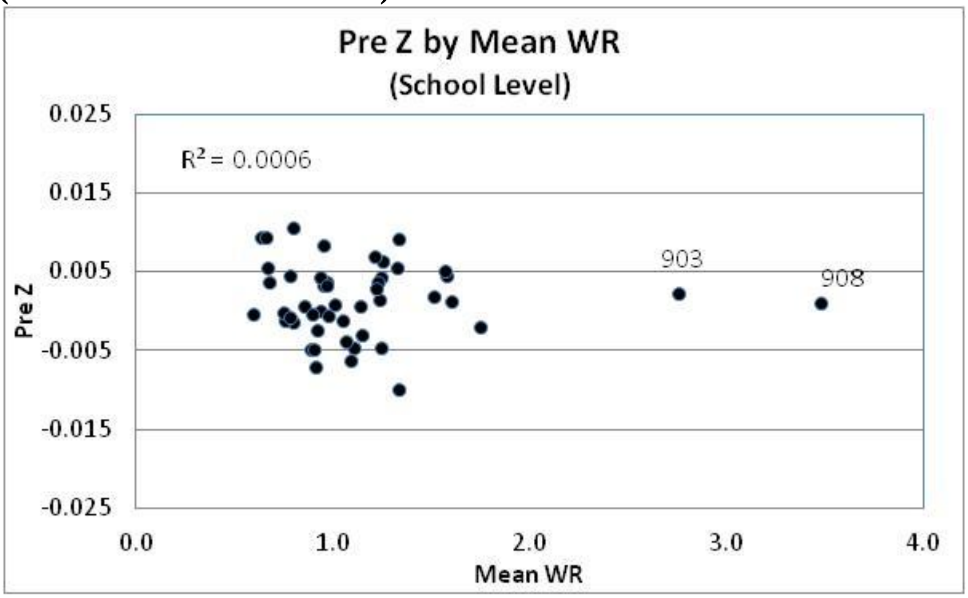


**Figure 12. Pre and Post Erasure Z by Mean Wrong to Right at the School Level (Mathematics Content Area)**



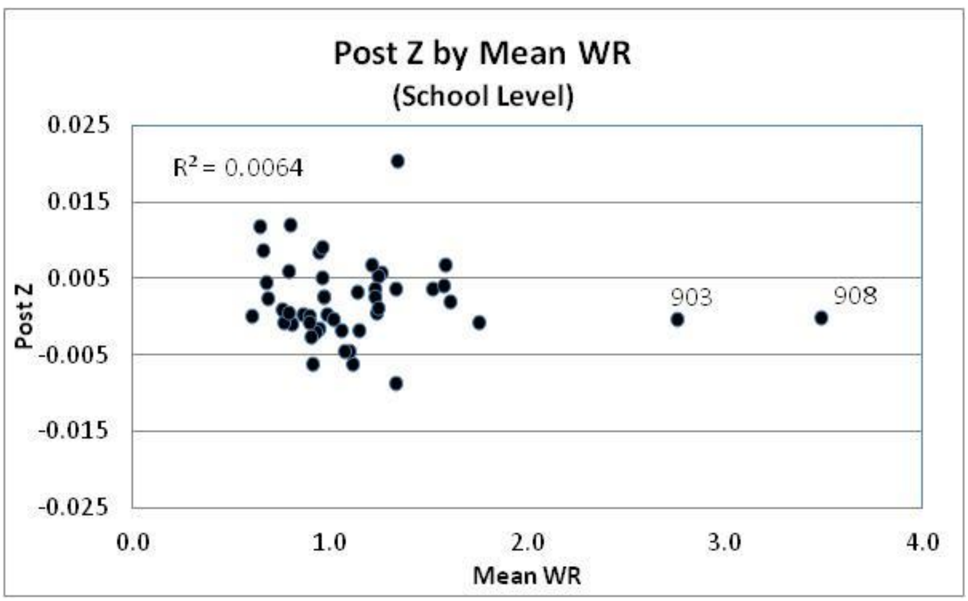



**Figure 13. Post Erasure Z by Pre Erasure Z at the School Level (Mathematics Content Area)**

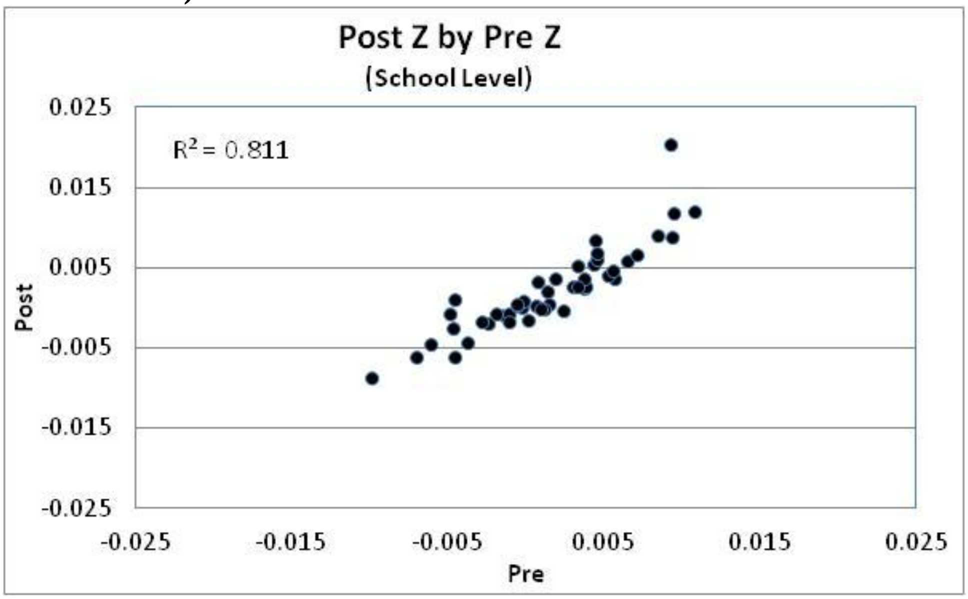


**Figure 14. Mean Wrong to Right by Total Erasures (Reading Content Area)**

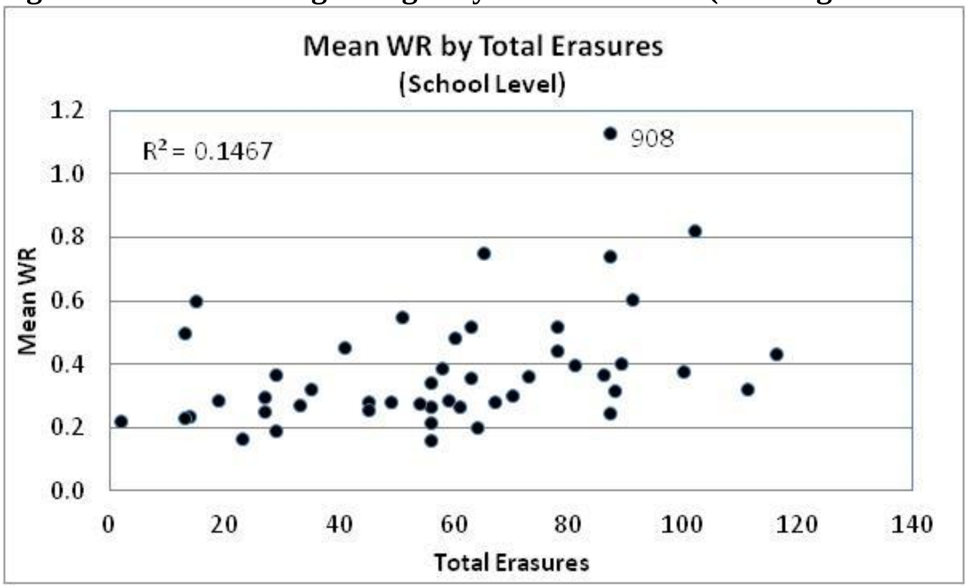


**Figure 15. Mean Post Erasure School Reading Achievement by Mean Pre Erasure School Reading Achievement**

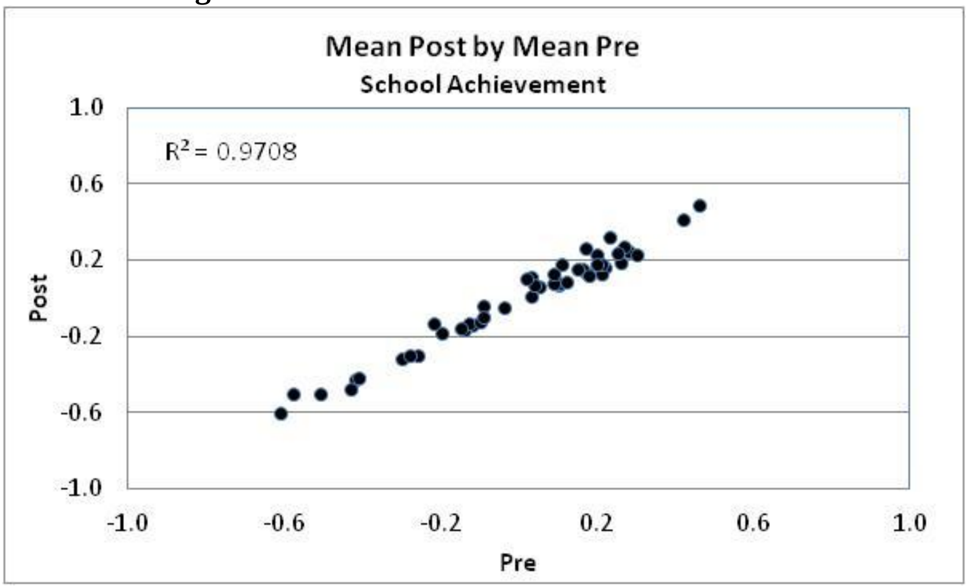


**Figure 16. Pre/Post Erasure Indicator by School (Reading Content Area)**

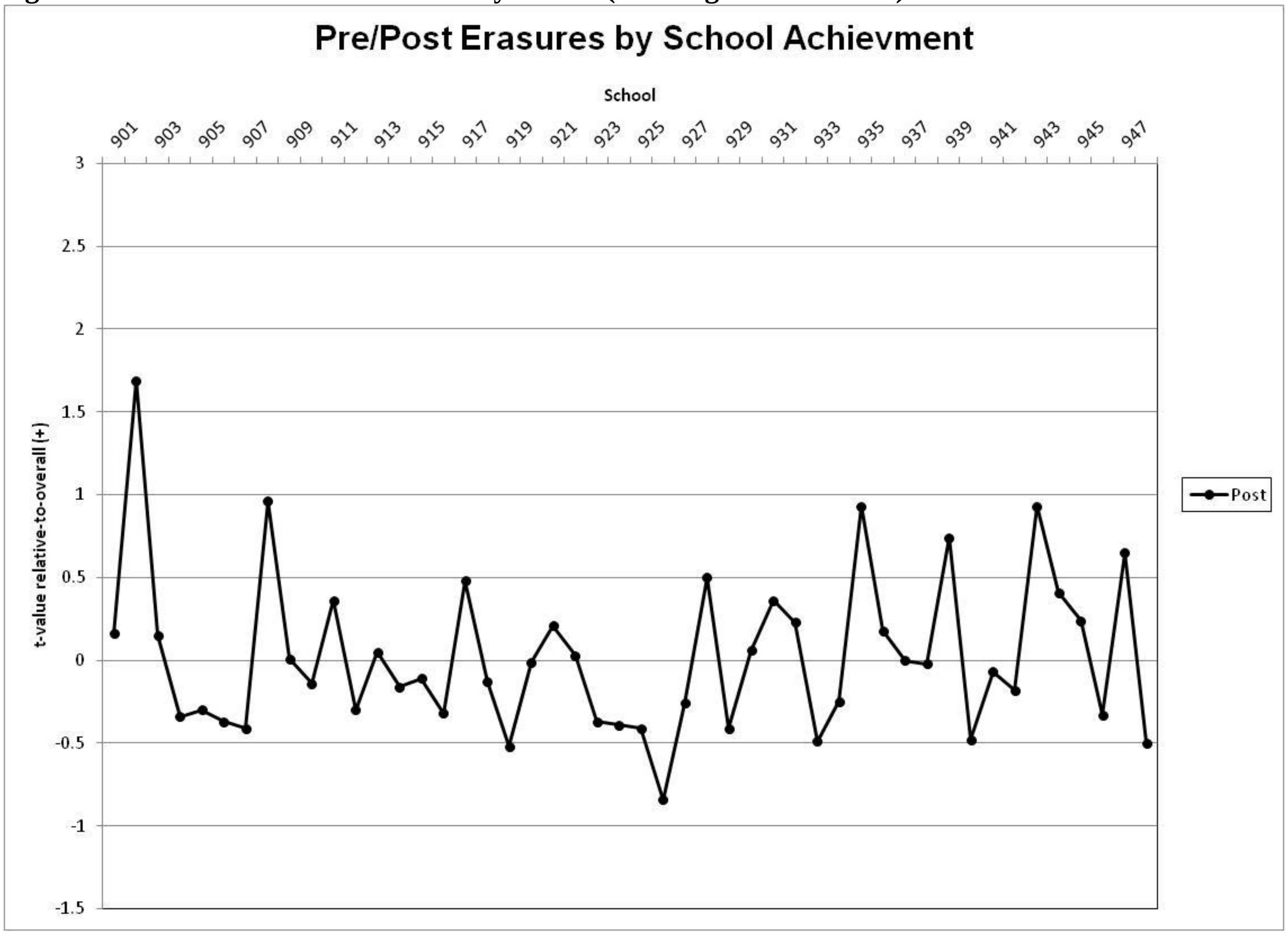
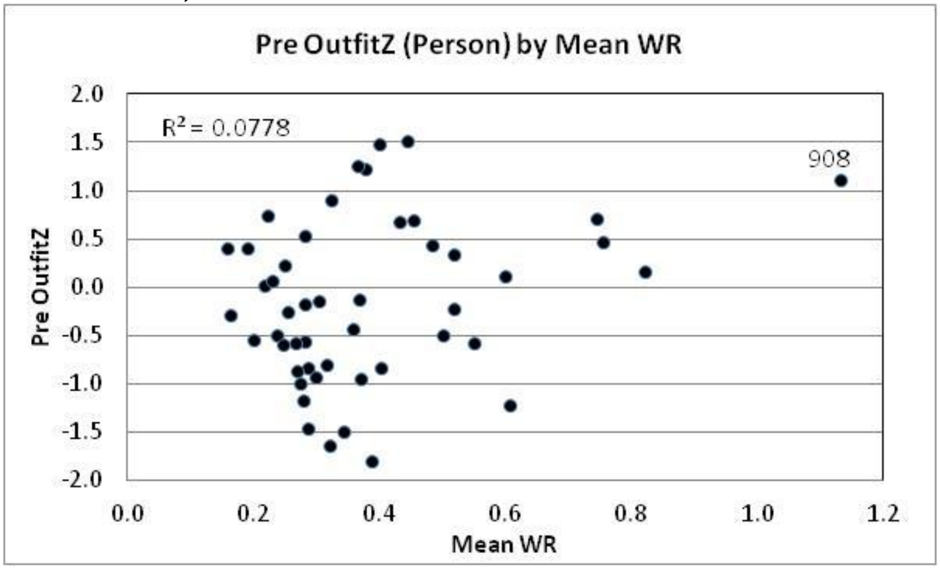


**Figure 17. Pre and Post Erasure Outfit Z by Mean Wrong to Right (Reading Content Area)**



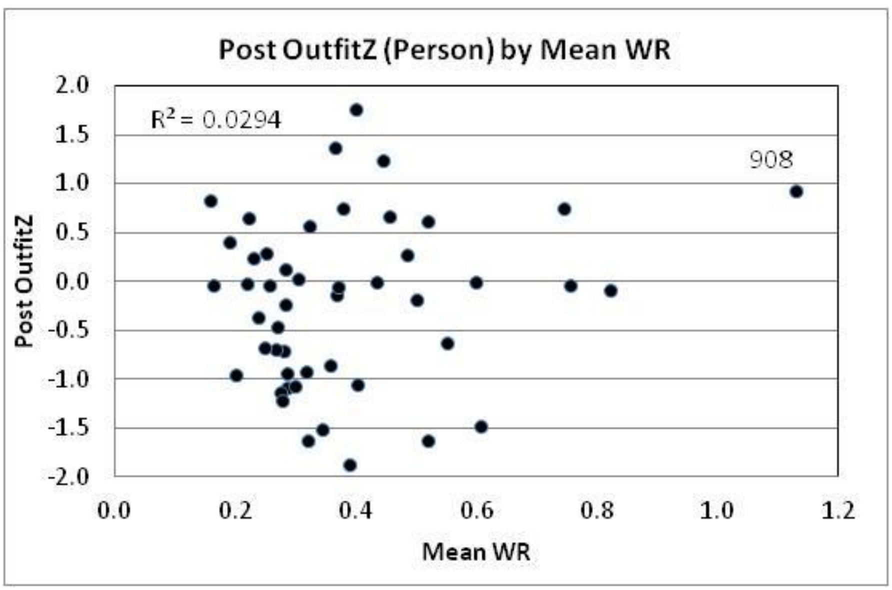
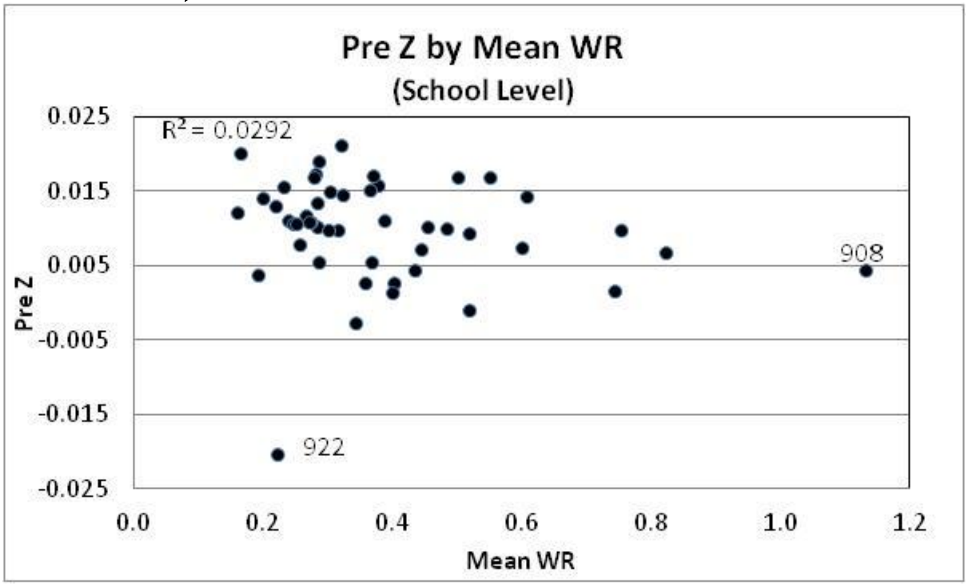



**Figure 18. Pre and Post Z by Mean Wrong to Right at the School Level (Reading Content Area)**



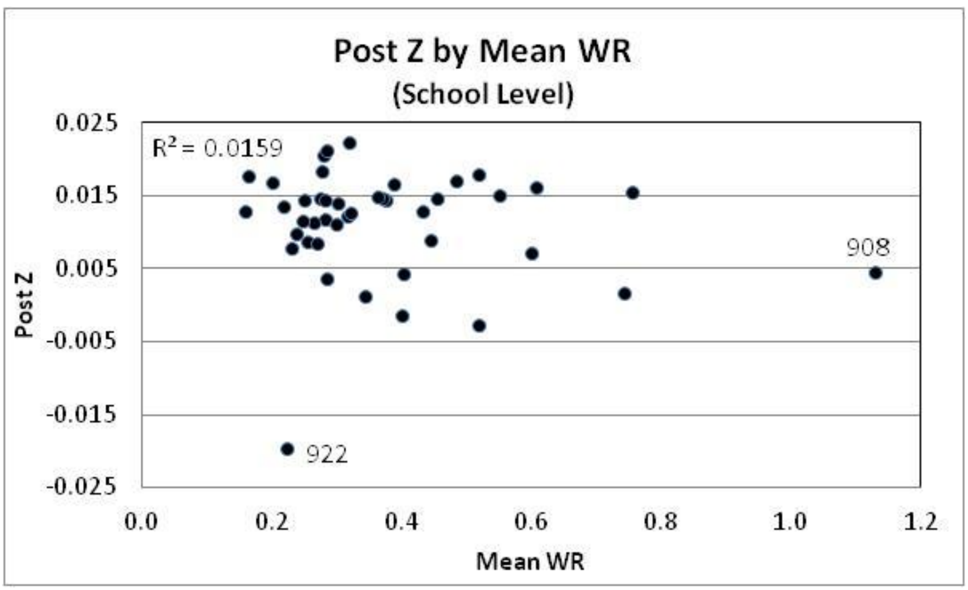



**Figure 19. Post Erasure Z by Pre Erasure Z at the School Level (Reading Content Area)**

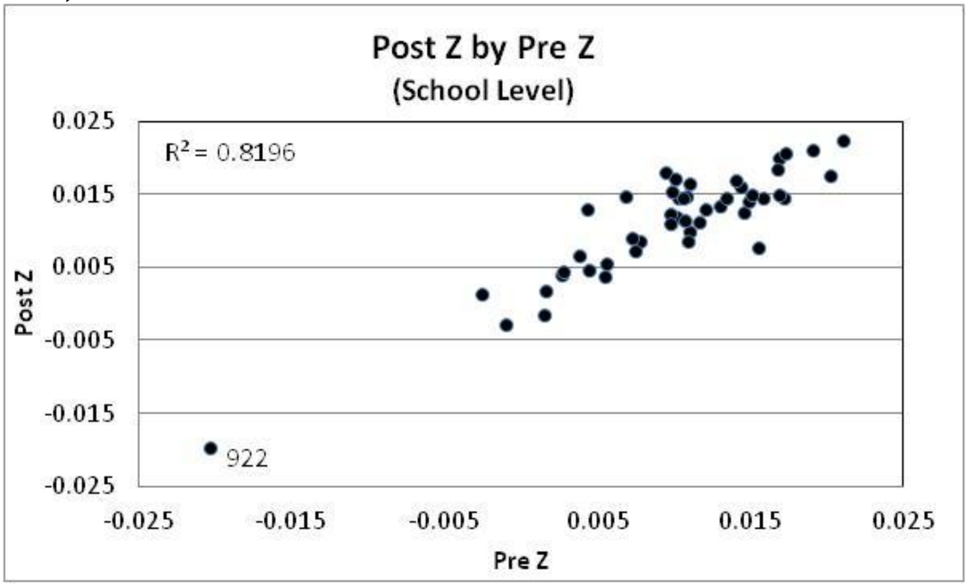
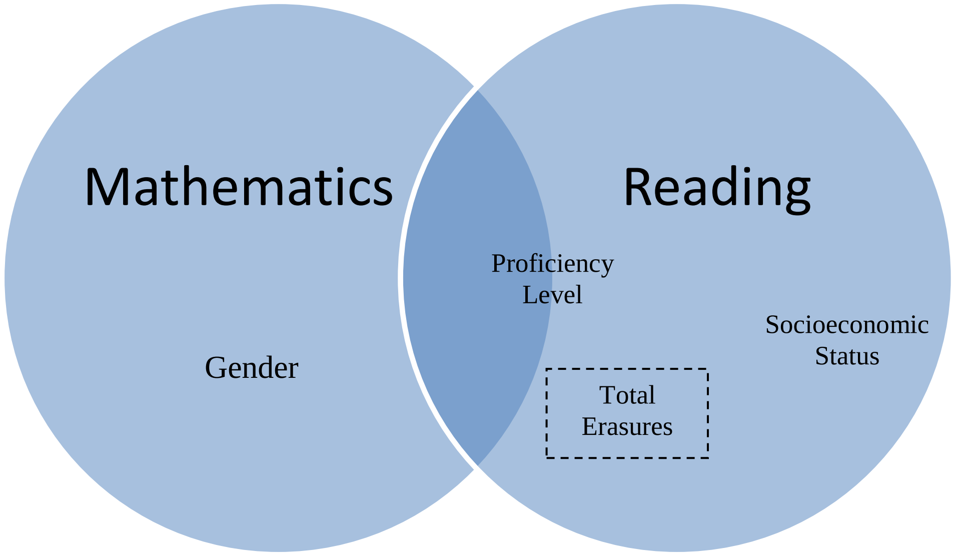


**Figure 20. Relationship of Covariate for HGLM by Content Area**



**Appendix A: IRB Determination Letter**

EMORY UNIVERSITY | Institutional Review Board

February 22, 2011

Aminah Perkins
Emory University
Division of Educational Studies
1784 North Decatur Road, Ste 240
Atlanta, GA 30322

RE: Determination: No IRB Review Required
IRB00049272 – Using Person Fit Analysis to Examine Erasure Data
PI: Aminah Perkins

Dear Ms. Perkins

Thank you for requesting a determination from our office about the above-referenced project. Based on our review of the materials you provided, we have determined that it does not require IRB review because it does not meet the definition(s) of "research" involving "human subjects" or the definition of "clinical investigation" as set forth in Emory policies and procedures and federal rules, if applicable. Specifically, in this project, you will be reviewing de-identified data obtained from the [redacted] Department of Education. With the data set you receive, you will be unable to determine any individuals' identities.

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

Thank you for consulting the IRB.

Sincerely,

Tom Penna
IRB Analyst Assistant
This letter has been digitally signed

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