Summer 2020

A Rasch Investigation of Type I Error Rates and Power Associated With Item Fit Statistics Under Large-Scale Testing Situations

Yin Burgess

A RASCH INVESTIGATION OF TYPE I ERROR RATES AND POWER ASSOCIATED WITH ITEM FIT STATISTICS UNDER LARGE-SCALE TESTING SITUATIONS

by

Yin Burgess

Bachelor of Arts
Guangzhou University, 1997

Master of Arts
Concordia University, 2004

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2020

Accepted by:

Christine DiStefano, Major Professor

Tammiee Dickenson, Committee Member

Richard Smith, Committee Member

Jin Liu, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

DEDICATION

I would like to dedicate this endeavor to my loving and supportive husband, without whose patience this would not be possible. I would also like to dedicate this work to my lovely children, who have challenged my endurance and patience to make me a better individual. Most importantly, I can never express enough indebtment and gratitude to my parents, who have unconditionally love and support me along the way to pursue my dreams and achieve my goals.

ACKNOWLEDGEMENTS

When I decided to start this journey, I was unsure where it would lead me. I definitely did not expect that it would transform me personally, academically, and professionally. However, this journey would not have been possible without the mentoring, support, and encouragement from my advisor, Dr. Christine DiStefano, who held my hand when my fledge was new yet patiently waited and cheered when my growth was steady. I could never forget the nurturing from Dr. Tammiee Dickenson and Dr. Ashlee Lewis, who have provided me with countless opportunities to apply the knowledge and skills I have acquired and allowed me great flexibility in balancing study, work and family. A very special thank you to Dr. Richard Smith, who selflessly shared his expertise and insight on a topic I am passionate yet knew so little about. I can never adequately express my gratitude to Dr. Robert Johnson, who lead me into the world of assessment and showed me kindness and care in academia. Thank you, Dr. Jin Liu, a special friend, colleague, and fellow researcher with whom I had numerous conversations about statistical procedures, research ideas, and future plans. There is also a place in my heart for my friends and colleagues at the REM Center and the Policy Center, who have witnessed and partaken this journey with me. I am ready for the next adventure.

ABSTRACT

This simulation study investigated the accuracy of the mean square and the standardized values of the item INFIT and OUTFIT statistics (i.e., based on total item fit) in Rasch dichotomous model under large-scale testing situations. It also examined their associated Type I error rates to determine how the rule-of-thumb critical values perform in detecting item misfit. Furthermore, simulated systematic measurement disturbances were used to test the power (i.e., the hit rates of true positive cases, true positive rates) and the false positive rates (i.e., Type I error rates) of the obtained values through between-item fit indices in identifying poor-fitting items. A total of four sample sizes (i.e., 5,000, 10,000, 25,000, and 50,000 test-taking students) and three test length (i.e., 30, 50, and 70 multiple-choice items) conditions were simulated to study how these statistics perform. Additionally, different percentages of items (i.e., 4%, 10%, 20% and 40%) with moderate to large uniform DIF (i.e., 0.35, 0.45, 0.55, and 0.65 logit units) were designed to test the power as well as the Type I error rates. The measurement disturbances were simulated between two balanced groups with "C" category DIF as defined by the ETS guidelines. Results suggested that ±2.0 for standardized values may be recommended for large-scale testing situations. Furthermore, it was found that the DIF item detection procedure currently used by Winsteps® is based on logistic regression, which, is sensitive to sample size and resulted in large numbers of items to be incorrectly identified with DIF.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

P_i , p	Probability of correctly answering an item *i* (where i = 1, …, *L*).
θ	Latent trait.
e	Base of the natural logarithm.
D	Constant. Usually 1.7 is used as a scale factor to covert the logistic distribution to a normal distribution.
a_i	Discrimination parameter for each item. It is the slope of a logistic curve.
b_i	Difficulty parameter for each item. It is the intercept of a logistic curve.
c_i	(Pseudo) guessing parameter for each item. It is the lower asymptote of a logistic curve and is restricted to positive values.
p_{ni} or P_{ni}	Probability of an examinee *n* (where n = 1, …, *N*) correctly answering an item *i* (where i = 1, …, *L*).
β_n	Ability estimate of examinee *n*.
δ_i	Item difficulty estimate of item *i*.
a	Discrimination parameter as a constant.
x_{ni}	Observed response of item *i* from examinee *n*.
y_{ni}	Observed response of item *i* from examinee *n* after *t*-transformation.
L	Test length.
N	Normal distribution.
U	Uniform distribution.
N_{R0K}	Number of examinees in the reference group who answer incorrectly.
N_{R1K}	Number of examinees in the reference group who answer correctly.

N_{F0K} Number of examinees in the focal group who answer incorrectly.

N_{F1K} Number of examinees in the focal group who answer correctly.

N_K Total number of examinees.

N Total sample size.

J Number of subpopulations.

N_j Number of examinees in each subpopulation.

d_{i1} Difficulty of item i in the calibration based on subpopulation 1.

d_{i2} Difficulty of item i in the calibration based on subpopulation 2.

s_{i1} Standard error of estimate for d_{i1}.

s_{i2} Standard error of estimate for d_{i2}.

LIST OF ABBREVIATIONS

ACT® ..... American College Testing

AERA..... American Educational Research Association

APA..... American Psychological Association

CAT.....Computer-Adaptive Testing

CI.....Confidence Interval

CTT .....Classical Test Theory

DIF ..... Differential Item Functioning

ELA.....English Language Arts

ELL ..... English Language Learning

EOCEP.....End-of-Course Examination Program

ESL .....English as a Second Language

ESSA.....“Every Student Succeeds” Act

ETS ..... Educational Testing Service

GRE® ..... Graduate Record Examinations

HSAP .....High School Assessment Program

ICC..... Item Characteristic Curve

INFIT ..... Information-Weighted Fit

IRT ..... Item Response Theory

MAP®.....Measures of Academic Progress

MAX .....Maximum

MH ..... Mantel-Haenszel

MIN ..... Minimum

MNSQ, MS ..... Mean Square

NAEP ..... National Assessment of Educational Progress

NCES ..... National Center for Education Statistics

NCLB ..... "No Child Left Behind" Act

NCME ..... National Council on Measurement in Education

NOSI-E ..... Nature of Science Instrument-Elementary

OECD ..... Organization for Economic Cooperation and Development

OUTFIT ..... Unweighted Fit

PCA ..... Principle Component Analysis

PISA ..... Programme for International Student Assessment

SAT ..... Scholastic Assessment Test

SCAAP ..... South Carolina Arts Assessment Program

SCDE ..... South Carolina Department of Education

SCPASS ..... South Carolina Palmetto Assessment of State Standards

SC READY ..... South Carolina College-and-Career-Ready Assessments

SD ..... Standard Deviation

TOEFL ..... Test of English as a Foreign Language

TST ..... True Score Theory

UB ..... Unweighted Between-Item Fit

UT ..... Unweighted

WB ..... Weighted Between-Item Fit

WT ..... Weighted

ZSTD ..... Standardized Z

CHAPTER 1

INTRODUCTION

Tests are an accepted part of our society. For example, preschoolers demonstrate what they know about colors and shapes before admittance into a Montessori program. High school graduates show their knowledge on specific content areas before they are accepted to a college. Doctors and nurses are required to showcase their medical knowledge and ability to work in the medical field prior to being granted a license to practice. Novice drivers must pass a knowledge portion and a road test at their local Department of Motor Vehicles office to obtain the permission to drive independently, etc. These scenarios remind us that testing is prevalent and commonplace in our modern society, and probably all citizens have experienced numerous testing situations before they enter the workforce.

Starting around the 1830s, formal assessments in the U.S. have gradually been standardized to reflect two fundamental beliefs about educational opportunities in the nation: fairness and efficiency (U.S. Congress Office of Technology Assessment, 1992). It is hoped that children are offered educational opportunities and services that are similar to those of other children, regardless of their background and previous experiences. Fair testing brings forth the attempt to minimize subjectivity; thus, test developers stride to develop items and tests that measure students' knowledge and performance as objectively as possible. According to Renée Grinnell (2018), objective tests are psychological instruments that measure "a person's characteristics independently of the taker's bias or

personal beliefs". Furthermore, "the results of such tests are generally more reliable than subjective (self-report) or projective (expose unconscious perceptions) models" ("PsychCentral," n.d.). In assessment, an objective test consists of questions "that can be quickly and unambiguously scored by anyone with an answer key", and this minimizes "subjective judgements by both the person taking the test and the person scoring it" ("Objective Test," n.d.). For instance, multiple-choice items are considered objective, as test takers typically receive the same raw score regardless of who scores the test. Other advantages of objective tests include time efficiency and cost effectiveness, as rater training is minimal and the rating process is not as time-consuming. All of these are attractive features, especially in large-scale testing situations in which the number of test takers can easily reach tens of thousands, and multiple testing sites are commonplace.

Traditionally, test takers obtain a sum score for the number of correct responses they have provided. This raw score indicates how well a test taker performs on a latent trait (i.e., an underlying construct, Bond & Fox, 2012; or a psychological attribute, Crocker & Algina, 2008) that the test intends to measure. Typically, a higher score suggests better performance on the latent trait, such as math achievement, medical knowledge, language ability, etc. For instance, in a Test of English as a Foreign Language (TOEFL; ETS, 1964), a score of 600 suggests that the English as a Second Language (ESL) learner has better mastery of the English language than another learner with a score of 380. However, the conclusion is not as clear-cut between two examinees' scoring, for example, 510 and 530 once measurement error is taken into consideration. Briefly, measurement error is considered as disturbances that are introduced by construct-irrelevant variables that may be due to the test-taker themselves (e.g., dyslexia or anxiety)

or more random/systematic occurrences (e.g., differences in testing occasions, differences in question formats, differences in social-economic background). In other words, the observed scores are not deemed to be a 1-to-1 representation of a latent variable when measurement error is considered (Bond & Fox, 2012).

When scoring tests, the field of measurement has largely moved beyond reporting the number correct (i.e., a "total" score) to use more complex methods to reflect a test taker's performance. Modern test theory methods like the family of Rasch measurement models (Rasch, 1960) and Item Response Theory (IRT; Birnbaum, 1968) examine other major contributing factors like item characteristics and rater variability in addition to providing a summative score. Thus, these methods improve precision in the testing situation, and ultimately, in score interpretation, which may lead to better decision making (i.e., validity associated with the scores).

Rasch Models

As a modern measurement framework to improve validity, Rasch measurement has been popular in state assessment programs. Rasch models assert the position that an instrument needs to be invariant in order to measure a construct. In other words, for items, one unit of the measurement scale (i.e., relative difficulties) should remain stable across substantially different subpopulations and regardless of what items are encountered (Bond & Fox, 2007). Furthermore, this psychometric technique was developed to improve the precision when practitioners develop instruments, monitor quality, as well as compute responses (Boone, 2016).

Rasch models belong to a family of psychometric models which share a common trait, with the existence of sufficient statistics (i.e., a statistic that summarizes all the

information in a sample about an unknown parameter without losing any information about this chosen parameter; Fisher, 1922) to estimate person and item parameters. Rasch models can utilize dichotomous and/or ordinal information as input for analyses. The underlying requirement is that data must fit the model, suggesting that data need to correspond with the underlying measurement paradigm. Misfitting items (or people) are problematic, as data, which do not conform to the underlying model, jeopardize the validity associated with the obtained scores.

To assess how well the data fit the model, Rasch methodology includes a variety of indices and tests for support. While a variety of methods have been developed to assess aspects such as unidimensionality and reliability, the most commonly used methods, however, focus on assessment of item and person characteristics. These methods use fit indices (called "fit statistics") to evaluate the performance of responses, grouped by either persons or items. The person fit statistics evaluate if responses from a person demonstrates the stochastic (i.e., probabilistic) structure of the Rasch model, in other words, if a person demonstrates an unexpected pattern in answers to the set of test items. For instance, a high-ability person provided an incorrect answer to an easy item; or alternatively, a person with low ability correctly answered a difficult question. The fit statistic would note this person as potentially misfitting as these occurrences occur more frequently in a response pattern. The item fit statistics evaluate if the responses to the item function as intended and thus, can separate individuals based on their ability levels. For example, a very difficult item was correctly endorsed by many low-ability test-takers while an easy item was incorrectly answered by many individuals with high-ability levels. The fit statistic would indicate these two items as potentially misfitting. As both

person and item misfit can be assessed, the focus here is on item fit because researchers are generally more concerned with item fit for a variety of validation purposes like selecting well-performing items to create a fair test.

Assessing item fit with the Rasch model. The main item fit statistic used to identify items that do not fit the Rasch model (i.e., misfitting items), is a Pearsonian chi-square initially proposed by Wright and Panchapakesan (1969). The chi-square statistic was based on the differences between the observed and the expected score for a group of examinees that have obtained the same raw score (Smith, Schumacker, & Bush, 1998). Subsequently, fit statistics were developed on the basis of the item residuals, which are the difference between the observed and the expected responses (Smith et al., 1998).

The fit statistic focuses on two versions: the information-weighted fit statistic (known as the "INFIT" statistic) and the unweighted fit statistic (known as the "OUTFIT" statistic). The INFIT statistic is a weighted index that gives more "weight" to the person whose measures are closer to the item's difficulty value. The OUTFIT statistic is unweighted, meaning it gives equal weight to the person regardless of their ability level, and is thus, more sensitive to outlying person scores. The OUTFIT statistic, being more sensitive to unexpected observations, is affected by random disturbances like guessing or carelessness (Smith, Schumacker, & Bush, 2000). Therefore, in comparison to OUTFIT, INFIT problems are more difficult to detect and manage, and they pose more threat to measurement than OUTFIT problems (Linacre, 2002b).

Problems with assessing model misfit. Both the INFIT and OUTFIT indices can be expressed in mean square (MNSQ) values and the standardized (ZSTD) values. When researchers use the INFIT and OUTFIT item fit statistics to investigate responses

to items, both values present a dilemma. Mean square values take into consideration the sample size. These values will illustrate good model-data fit for an item simply by increasing the sample size. Most (if not all) items will 'appear to' fit well in a situation when the sample size is large enough. In this large sample situation, if ZSTD values may be used instead, however, with ZSTD, even the smallest amount of misfit becomes significant. Therefore, most items will likely be rejected (Bond & Fox, 2012).

The dilemma concerning what variant of the item fit statistic (MNSQ or ZSTD) to use may have important consequences in many situations that administer a longer (e.g., 50 items) test to many examinees. For example, in a typical state-wide testing situation when the number of examinees is large, using the MNSQ values to evaluate item fit may result in misfitting items not being identified as they will 'appear' to fit simply due to the large sample size. By the same token, using the ZSTD values may over-identify items as majority of the items will have 'significant' amount of misfit due to the large sample size. Unless corrected for sample size (Smith, 1982), using MNSQ would lead to misidentification of items in testing when problematic items are deemed good, or good items are specified as misfitting compared to the use of ZSTD values. Therefore, in large-scale testing situations, what fit index should be used to examine item fit? What Type I error rate (i.e., type of errors researchers make when a false alternative hypothesis is concluded to be true, Sheskin, 2007; in this study, it is defined as the percentage of cases when well-performing items are incorrectly categorized using specific cut-off values, or false positive rates) should be used for the approximate critical values? Furthermore, what is the power (i.e., the percentage of cases when misfitting items are

correctly identified using specific cut-off values, true positive rates) of the approximate critical values in detecting measurement disturbances?

Statement of the Problem

Considering that many decisions made from state-wide assessments are high-stakes and bring major consequences like retention, or graduation, it is important for assessments to use psychometrically sound (i.e., high quality) items. Detecting malfunctioning items for future review or even exclusion is the initial step towards safeguarding fairness. In the Rasch framework, various indices have been proposed and used to identify misfitting items; however, it remains controversial as to which index is more accurate under situations that involve different numbers of test takers and different lengths of tests. Furthermore, it is questionable if the general guidelines to determine misfit of items function well in different testing situations. In other words, how accurate are the suggested indices in identifying item misfit, especially in items that function substantially differently for different subgroups?

This study specifically focuses on large-scale testing situations such as state-wide assessment programs in which sample size is very large and the length of the test is medium to long (e.g., 50 or more items). Large sample sizes may result in under-identifying problematic items if MNSQ values are used and in over-identifying misfitting items if ZSTD values are selected (Bond & Fox, 2012). Therefore, it is necessary to investigate which Rasch-based fit indices provide more accurate information about individual items, how accurate the information is, and the ability (known statistically as the power) of these indices to detect measurement disturbances (i.e., interfering

conditions that affect the process of measurement of a latent trait; Schumacker, Mount, & Marcoulides, 2005).

To investigate this issue, a simulation study was conducted. The present study has attempted to provide some guidelines on item fit indices for large-scale testing situations, and it is hoped that the information can help psychometricians and test developers make better decisions about the item selection for future test administrations in order to provide fair testing opportunities.

Research Questions

As the Rasch methodology is widely used by state agencies to analyze standardized test results, it is imperative that information on the test items is accurate to inform stakeholders to make appropriate decisions. Since typical state-wide assessment programs involve large-scale testing situations with long tests, the purpose of the study is to investigate the following research questions:

1. Which value of the total item fit statistic is more sensitive to the large sample size (i.e., 5,000, 10,000, 25,000, 50,000) for an assessment of 30, 50, or 70 items, MNSQ or ZSTD values for INFIT and OUTFIT statistics?
2. What Type I error rate should be used for the approximate critical values for each of the above conditions?
3. What is the power of the obtained critical values to detect simulated measurement disturbances?

CHAPTER 2

LITERATURE REVIEW

Assessment plays an important role in our modern daily life, and it is especially relevant in the field of education. Assessment provides stakeholders with information on student achievement and growth, informs teachers about instructional decisions and focus, and guides politicians on the making of educational policies. Assessment quality, and, particularly item quality, is essential to improve accuracy of the assessment results as well as validity of the conclusions drawn. Information regarding item performance helps test developers identify items that may not function as intended on the instrument and need to be revised or even excluded from the assessment.

This chapter will briefly describe the history of large-scale assessment, its role in education as well as how information from large-scale tests has been used. Then, after a short discussion on the traditional measurement approach Classical Test Theory (CTT; Spearman, 1907, 1913, also referred to as True Score Theory, TST; Lord, 1964) for computing test results, it will elaborate on commonly used modern measurement frameworks of Item Response Theory and Rasch measurement. After a detailed description of Rasch measurement, the chapter will showcase how Rasch models contribute to large-scale assessment programs and how measurement findings obtained from Rasch models have been utilized to make educational decisions. Next, the chapter will discuss selected Rasch-based fit indices and how these indices have been utilized for

measurement purposes. The final section of this chapter will focus on item fit statistics, a widely used approach by practitioners to identify misfitting items. After a brief explanation on how measurement disturbances may impact item performance, a detailed review of recent studies involving Rasch-based item fit statistic is discussed.

Large-scale Testing

Large-scale testing has become commonplace in our modern society, and almost all citizens have encountered such testing situations for important decisions, such as licensure, promotion, graduation, or certification. The process has become standardized in an attempt to provide consistent and reliable performance information from the test results.

Brief history of testing. Systematic nation-wide standardized testing was first conducted in the Sui Dynasty (581-619) in ancient China, with the purpose to select candidates for government positions ("Imperial Examination in China", n.d.). Formalized during the Tang Dynasty (618-907), annual tests were administered in content areas like literacy composition, arithmetic, and legislation to fill official and officer vacancies at various national and provincial levels ("Imperial Examination in China", n.d.).

In the early 19th century, the idea of standardized testing was advocated by Britain's counsel in Guangzhou, China and introduced to Europe. The British Empire first adopted testing practices for managers to make hiring and promotional decisions on company employees in order to prevent corruption and favoritism ("Imperial Examination", n.d.). Testing then spread throughout the British Commonwealth, Europe, and to America.

Serving a different purpose in America, standardized aptitude quizzes named Army Mental Tests were administered during World War I to assign jobs to US service personnel ("History of Standardized Testing in the United States", n.d.). These tests systematically evaluated soldiers' intellectual and emotional well-being. Named the Army Alpha, the examination measured "verbal ability, numerical ability, ability to follow directions, and knowledge of information" and was administered to nearly two million servicemen ("History of Military Testing", n.d.). Its nonverbal equivalent, the Army Beta, was subsequently developed for illiterate and foreign soldiers (Pintner, 1921).

Testing in the school setting. In the US, large-scale educational testing began state-wide in the 1970s and in the 1980s assessment became nation-wide in response to accountability demands in public schools which began gaining momentum in the 1960s (Stiggins, 2002). While legislators attempt to use assessment outcomes to guide the decision-making process of education policies, educational laws exert substantial impact on the shape of K-12 public education in the United States. Reauthorization of a major federal mandate, the Elementary and Secondary Education Act of 1965, known as the "No Child Left Behind" Act (NCLB, 2002), has driven education policy since its passage in 2002 (Dappen, Isernhagen, & Anderson, 2008) and brought assessment to the forefront in education. NCLB requires states to administer annual assessments in reading, math, and science, and uses the student academic achievement information to provide evidence for school accountability. Despite its problems and critics, the law has resulted in an increase in state-wide standardized testing, higher demand for performance standards, as well as more serious consequences for students, educators, and schools (DePascale,

2003). In December 2015, the NCLB law was replaced by the "Every Student Succeeds" Act (ESSA, 2015), which modified the previous legislation but maintained provisions that were related to periodic standardized testing ("Every Student Succeeds Act", n.d.).

Large-scale assessment is defined as tests that are "administered to large numbers of students, such as those in a district or state" (Montana Office of Public Instruction, as cited in DePascale, 2003, p. 3). Generally, these assessment programs test students' ability in content areas such as language arts, math, social studies, science, and history (DePascale, 2003). They are usually administered to all students at various grade levels in elementary, middle, and high schools (DePascale, 2003). The National Assessment of Educational Progress (NAEP; National Center for Education Statistics, 1969), American College Testing (ACT®; ACT Inc, 1959), Scholastic Assessment Test (SAT; College Board, 1993), and Measures of Academic Progress (MAP®; Northwest Evaluation Association, 2013) are some of the nation-wide tests students typically encounter in addition to state mandated testing. Furthermore, these large-scale assessments have multiple objectives, including improving instruction (Landau, 1998; Popham, 2001), measuring students' academic performance ("Measured Progress", 2003, as cited in DePascale, 2003), determining education quality (Popham, 2001), and informing policy decisions (Wu, 2010).

Test format. As objectives of large-scale testing have changed over time, so have testing conditions and the test format. Large-scale standardized testing has evolved from the paper-and-pencil format to the electronic computer-administered tests with fixed (i.e., an identical set of items to all examinees) or scrambled items (i.e., a set of items with different sequences or random ordering). In addition, the computer-adaptive testing

(CAT) has gain popularity. Initially developed by Reckase (1974), an interactive computer program was an early example of the CAT systems for tailored testing (Linacre, 2000). As one of the most important developments of the 1990s, CAT systems adjust/adapt the test items that are administered on an assessment to an individual test taker's ability level (Parhizgar, 2012). For instance, if a test taker responds correctly to an item, the subsequent question will be slightly more difficult/challenging. Vice versa, the next test item will be slightly easier/less challenging when a test taker fails a previous item. In other words, items are selected based on the current test-taker's estimated ability level (Linacre, 2000). Targeting item difficulty on person ability increases the accuracy of the examinee's ability by lowering the standard error of measurement. When the required accuracy for computing an examinee's ability is reached, the test concludes (Linacre, 2000). Thus, different test takers will, most likely, not encounter identical sets of test items.

Item format has also developed from the traditional multiple-choice, closed-ended, to more task-based performance methods such as essay composition, reading report, oral presentation, drawing, etc. The benefits for using performance as assessment format are abundant. According to Johnson, Penny, and Gordon (2009), performance assessment is authentic and reflects real world situations in the field. It also requires complex cognitive skills and strategies like analyzing and synthesizing to complete a task. Furthermore, performance assessment demonstrates in-depth content knowledge and skills in the field (Messick, 1996). Nevertheless, performance-based tasks require an extended period of time to develop, to complete, and to score. They usually involve

substantial costs in test administration and scoring, and these tasks do not cover as much content in comparison to multiple-choice items (Johnson et al., 2009).

Use of test information. As large-scale assessment plays an increasingly significant role as an accountability tool, results of student academic performance have also been used for various educational purposes. For instance, state agencies use test results to evaluate education quality and allocate funding accordingly; districts use information from tests to identify areas of strength or for improvement and provide targeted support to each school; school teachers use report card information to guide instruction and adjust teaching strategies; parents and students use this information to obtain a comprehensive view of the child's academic performance. Very often, the decisions are high-stakes and involve situations such as student retention, promotion, or graduation (Wyse, 2011; Yen & Henderson, 2002). With time, the creation of technically sound assessments has gained great attention in order for stakeholders to face the intense professional and public scrutiny (Goodman & Hambleton, 2004).

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) explicitly state that "the higher the stakes associated with a given test use, the more important it is that test-based inferences are supported with strong evidence of technical quality" (p. 139). High-quality assessments, particularly high-quality assessment items with solid psychometric properties, produce accurate and consistent results, which lead to conclusions that will guide crucial decision making at the national or state levels.

Traditional and Modern Measurement Models Used to Evaluate Test Quality

As psychometrically solid tests are essential to providing reliable and valid information obtained from test results, researchers and test developers must focus on improving item quality. Commonly used methods to examine item quality include traditional approaches such as CTT as well as modern measurement methods such as IRT and Rasch.

Classical Test Theory (CTT) methods. Traditionally, a test's item-level information is obtained through the framework of CTT, which uses information from observed scores. In identifying problematic items (e.g., items that are correctly/incorrectly answered by the majority/all test takers), CTT has the disadvantage of producing sample-specific item parameters and interpretations limited to the group being tested (Alasuutari, Bickman, & Brannen, 2009). The dependencies of both person and item parameters on the test and examinee sample can limit the utility of these parameters in practical test development work (Hambleton & Jones, 1993).

Furthermore, under CTT, item calibration (i.e., creation of "a common scale on which all item parameter estimates can be expressed"; Crocker & Algina, 2008, p. 363) and person ability do not use a common scale (Perline, Wright, & Wainer, 1979; Rasch, 1960; Smith, E., 2004). Therefore, researchers cannot predict how an examinee would perform on an assessment, given his/her level of ability (Smith, E., 2004). Another limitation of CTT is that score interpretation and reporting differ if a norm-referenced assessment or a criterion-referenced assessment is presented (Smith, E., 2004). The former provides information on individual examinees' relative position regarding other examinees in the norm group (e.g., different gender, ethnicity, age groups, etc.) but,

proficiency level remains unknown. The latter illustrates examinees' level of proficiency in a particular area yet, an examinee's relative standing to the others is unclear. In sum, the traditional CTT framework limits the usage and interpretation of scores.

Nevertheless, CTT information is still usually reported as this framework has several advantages (DeVellis, 2006). One major advantage is that researchers and practitioners are typically familiar with its basic concepts because they serve as building blocks to modern measurement theories. Another is that major statistical packages (e.g., R, SAS®, SPSS) include procedures to routinely perform basic analyses, such as reliability calculation using coefficient alpha (Cronbach, 1951). These statistical packages are "widely available and relatively easy to use" (DeVellis, 2006, p. S57).

As educational researchers and practitioners strive to improve assessment validity, methods to examine various aspects of test development have evolved from the traditional CTT to include modern frameworks of Item Response Theory (IRT) and Rasch measurement models.

IRT models. Popular IRT models include one-parameter 1PL, two-parameter 2PL, and three-parameter 3PL models to estimate an examinee's ability level on a latent trait from their response pattern to an item. The three item parameters, denoted by *a*, *b*, and *c*, respectively represent the item discrimination, difficult, and (pseudo) guessing information. Each model can be expressed through a mathematical formula and uses item information from one, two, or three of these parameters. The 2PL and 3PL models are briefly explained first, followed by the Rasch model, which is mathematically equivalent to the 1PL model but philosophically distinct.

Focusing on the latent trait score instead, IRT models (e.g., two-parameter 2PL, three-parameter 3PL) assume that examinees' responses to a test item can be accounted for by a latent trait, and these responses provide information on how test takers from various ability levels perform on a particular test item (Crocker & Algina, 2008). Mathematically, there exists an underlying (mathematical) model which is used to estimate how an examinee from a particular ability level of the latent trait would respond to a test item (Crocker & Algina, 2008). This knowledge overcomes the limitation of the sample-specific information and allows comparison of examinee performance between different tests or items as well as comparison of item performance between different groups of test takers (Crocker & Algina, 2008).

Two-parameter model. Using logistic models, the 2PL model estimates two item parameters: discrimination (denoted by *a*-parameter, it is the slope of a logistic curve) and difficulty (denoted by *b*-parameter, it is the intercept of a logistic curve). Following Crocker and Algina (2008), the mathematical formula for this model can be expressed as:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \quad [1]$$

where θ symbolizes the latent trait, e is the base of the natural logarithm, and D represents a constant. Customarily, the constant $D = 1.7$ is used as a scale factor to convert the logistic distribution to a normal distribution (Wu, Tam, & Jen, 2016). In the 2PL model, the *a*- and *b*- parameters are specific to each item, and the probability of correctly responding to an item is estimated through these two parameters as well as an examinee's ability level on the latent trait.

In additional to the advantages of modern measurement methods over the traditional CTT approach, some of the 2PL model specific advantages include "providing

a closer statistical fit to the empirical data" and "more reliable ability estimates" (Baylor et al., 2011, p. 248). However, this model lacks sufficient statistics to estimate the slope parameter, consequently, this model requires large sample sizes for item calibration since more parameters need to be estimated (in comparison to only estimating item difficulty parameter). Furthermore, estimation of the discrimination parameter makes it more complicated to interpret and explain the item characteristic curves (ICC), which are curves that indicate the probability of correctly responding to an item given examinees' ability levels on the latent trait to suggest underlying performance of an examinee (Crocker & Algina, 2008). For instance, if Item 1 and Item 2 have different *a*-parameter values (i.e., item discrimination), they have different slopes. Their ICCs would intersect at some point, and examinees at different ability levels would have different probabilities of correctly responding to Item 1 and Item 2. This intersection of the ICCs violates the invariance property, a requirement for Rasch measurement models, that states an item needs to present the same probability of correctly responding regardless of an individual's ability level. In other words, the measurement unit needs to have the same distance for individuals at different ability levels.

Three-parameter model. In addition to the item discrimination and difficulty, the 3PL model also estimates the (pseudo) guessing for an item (denoted by *c*-parameter, it is the lower asymptote of a logistic curve and is restricted to positive values). Mathematically, this model can be expressed as:

$$P_i(\theta) = c_i \frac{(1 - c_i)\, e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \qquad [2]$$

where c_i represents the additional (pseudo) guessing parameter for each item (Crocker & Algina, 2008).

With the use of information from more item parameters, the model flexibility results in better model-data fit (CTB/McGraw Hill, 2008, as cited in LeBeau & McVay, 2017). Additionally, the 3PL model allows low ability test takers to obtain a greater than zero likelihood of correctly responding to an item through (pseudo) guessing (Crocker & Algina, 2008). However, information from more item parameters (i.e., *a*, *b*, *c*) introduces many problems. First, the 3PL score is not a sufficient statistic; meaning the total score does not contain all available information about the measurement of the latent trait. Furthermore, the score is no longer a logistic function; therefore, due to the addition of the estimation of the lower asymptote, it is problematic mathematically.

The 3PL model still requires large sample sizes to estimate item parameters like the 2PL model. Pelton (2002) pointed out that it is even more difficulty to accurately estimate the *c*-parameter than the *a*-parameter thus, increasing uncertainty in the obtained estimates. Furthermore, as the (pseudo) guessing parameter is calibrated based on weak or capable samples of examinees, the item parameter sets that are produced may be substantially different (Pelton, 2002). Therefore, researchers need to exercise care in interpreting perceived advantages in simulated 3PL studies on the accuracy of calibrated estimates (Pelton, 2002).

Rasch methodology. While the testing industry utilizes the two-parameter (i.e., 2PL) and the three-parameter (i.e., 3PL) models under the IRT framework (e.g., major educational tests like the Scholastic Aptitude Test [SAT] and Graduate Record Examinations [GRE®], An & Yung, 2014; Birnbaum, 1968; LeBeau & McVay, 2017), most state-wide assessment programs use Rasch methodology to provide results. The Rasch model is preferred by state-wide testing programs for many reasons. The primary

reason, however, might be contributed to the differences in the theoretical frameworks, in which the IRT framework models are more "concerned with accurate test scoring and development of test items" (An & Yung, 2014, p. 1). In contrast, the Rasch model is "developed to improve the precision" of item quality in order to "construct instruments, monitor instrument quality, and compute respondents' performances" (Boone, 2016, p. 1).

Rasch model. Named after the Danish mathematical statistician Georg Rasch, the Rasch model was developed during the late 1940s to the mid1950s from a probabilistic function that utilized the logistic distribution to model the probability of a correct response (Schumacker, 2004). In contrast to IRT, Rasch (1960) noted and subsequently proved that "only the item difficulty parameter can be consistently and sufficiently estimated" (Schumacker, 2004, p. 227) from dichotomous response data. Placing the item difficulty and the person ability on the same logit scale, the Rasch dichotomous model allows the item difficulty level to be independently calculated from the ability level of individual examinees. The mathematical formula for the basic Rasch model is defined as:

$$P_{ni} = \frac{e^{(\beta n - \delta_i)}}{1 + e^{(\beta n - \delta_i)}}, \qquad [3]$$

where P_{ni} is the probability of an examinee correctly answering an item, β_n is the ability estimate of examinee *n* (where n = 1, …, *N*), and δ_i is the item difficulty estimate of item *i* (where i = 1, …, *L*) (Schumacker, 2004, p. 229). Mathematically, this model is equivalent to the one-parameter (i.e., 1PL) IRT model with dichotomous data in which:

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}}, \qquad [4]$$

where Da equals to 1.7 multiples a constant a instead of a variable a_i as in the 2PL model (Crocker & Algina, 2008).

Depending on the response type, the Rasch family consists of models that accommodate: dichotomous responses (i.e., dichotomous Rasch model, Linacre & Wright, 1999; Rasch, 1960; Wright & Stone, 1979) and Likert rating scales (i.e., rating scale model, Andrich, 1978; Wright, Linacre, & Schultz, 1989; Wright & Masters, 1982; or the partial credit model, Linacre & Wright, 1999; Masters, 1982). If raters are involved in providing responses with a rating, more complex models can be employed, such as many-facet Rasch measurement (Linacre, 1994; Schumacker, 1999), linear logistic test model (Fischer, 1973), or paired-comparison model (Florin, 1999). Furthermore, more specialized Rasch models (i.e., multilevel measurement models, Adams & Wilson, 1996) have been developed to analyze data with more complex structures. These Rasch models provide researchers and practitioners with ample measurement tools to meet different analysis needs in various situations they have encountered.

Rasch model properties. The family of Rasch models demonstrate attractive measurement properties that include: 1) a sufficient statistic for measuring the location of both persons and items on the latent trait; 2) specific objectivity; and 3) local independence for items. The sufficient statistic is listed prior to the other two due to its primary importance. Under the Rasch model, this is observed as a logistic function that uses the person/item counts (i.e., the total number of people who correctly select the response to an item, or the total number of items that an individual correctly answers; Wright, 1989), and a common unit of measure logit (or log odds, the logarithm of the

odds of success as $p/(1-p)$ where p is the probability; Bond & Fox, 2012). This function creates a common reference scale for both person ability and item difficulty (Schumacker, 2004), allowing both person ability and item measure share a common scale. Statistically, this property allows person ability estimates to be freed from the distribution of the specific items encountered and also item difficulty estimates free from the specific people used for item calibration (Wright & Masters, 1982).

Specific objectivity suggests that the approach is independent of the sample of items or persons. This means that examinees would maintain their relative ranking in the group regardless of which items are selected to use in the assessment. For example, consider person A is more able in a certain latent trait than person B (i.e., higher level of the latent trait). Even though both persons A and B took assessment Form 1 the first time and Form 2 the second time, their scores from both forms should reflect that person A got a higher score than person B. When the analyses are sample-independent, it helps to build construct validity (i.e., inferences drawn from test scores to an intangible psychological trait that a single criterion cannot adequately represent or a universe of content cannot completely define; Crocker & Algina, 2008) evidence, as person and item measures are invariant.

Local independence is a requirement that suggests "the (conditional) distributions of the item scores are all independent of each other" (Lord & Novick, 1968, p.361). In the Rasch and IRT models, items in a test are required not to be related to each other and that the construct is the driving force behind the item. This suggests that a test taker's reply to one item in a test should not affect his/her answer to another item. In other words, the correlation of item residuals (i.e., differences between the observed and the

expected value for a respondent of a particular ability level on a test item; Linacre, 2002a) between two items is zero if the items are not correlated after the effect of the latent trait is removed (Baghaei, 2008).

Rasch model assumptions. As with other statistical methods, the family of Rasch models has assumptions that need to be met to ensure accuracy of estimation. There are three basic conditions (i.e., assumptions) that must be satisfied: 1) unidimensionality of the scale; 2) monotonicity of the scale; and 3) good model-data fit. These conditions suggest that the instrument should measure only one latent construct (i.e., unidimensionality), a higher latent score indicates a higher level of the latent trait (e.g., math ability), and the items fit the Rasch model (i.e., an item performs as intended, for instance, a difficult item being endorsed by examinees with high abilities instead of low abilities; Bond & Fox, 2012). Unidimensionality under the Rasch model may be tested by examining model residuals resulting from applying a one-factor principle component analysis (PCA) to the data to indicate presence of multiple dimensions. This analysis extracts data variance explained by the Rasch model/latent construct and determines if additional variance is present.

For the third model requirement, good model-data fit, global fit statistics such as Pearson chi-square, log-likelihood chi-square may be evaluated ("Global Rasch Fit Statistic", 2007) to illustrate how well data fit the Rasch model. Viewed as an ideal characteristic of measurement, invariance holds the key that provides evidence to construct validity, and it gives the Rasch models the advantage over traditional CTT methods and IRT models (Engelhard, 2013).

Rasch model information. The Rasch model provides information concerning how well the item/person (or both) responses conform to the proposed (underlying) model. Those items/persons that do not fit the Rasch model are thought to compromise the construct validity associated with the attained scores. Thus, misfitting items (or people) are problematic and need to be accurately identified as the Rasch model serves the purpose of constructing fundamental measures (Bond & Fox, 2012). Therefore, it is crucial that methods be used and statistical information be provided to suggest if analysts can confidently establish the relationship between ordering of a dependent variable (e.g., math achievement) and the joint effect of two or more independent variables (e.g., items to measure math and individuals' math ability). This is essential to justify the claims we make about the results obtained from this measurement scale with interval measurement properties (Bond & Fox, 2012). Information from the model provides evidence to researchers and practitioners the "extent to which the test may be said to measure a theoretical construct or trait" (Anastasi & Urbina, 1997, p. 126).

Use of the Rasch Model with Educational Decision Making. As a modern approach that evaluates an instrument at the individual item and/or examinee level, Rasch methodology has been widely employed in large-scale testing situations to examine psychometric properties of various educational measures. Primary uses are defined and illustrated with examples that relate to large-scale testing situations.

Scale development and validation. A major use of the Rasch model is scale development and validation. Moreover, it is a popular method to provide validity evidence to tests or scales. For instance, researchers have utilized Rasch model to scale latent scores in an international achievement program (Monseur & Adams, 2009), or to

obtain scaled scores on constructed-response items (Sykes, Yen, & Ito, 1996). Also, researchers have used Rasch-scaled scores to select benchmark writing samples (Popp, Ryan, Thompson, & Behrens, 2003) and to set standards for writing tasks in alignment with the European framework (Harsch & Rupp, 2011). Additionally, the Rasch partial credit model was employed to examine the effects of local item dependence on testlet calibration (Yan, 1997). In measuring elementary or middle level students, Rasch method was used to provide content, substantive, and generalizability evidence to develop the NOSI-E scale (Nature of Science Instrument-Elementary; Peoples, 2012; Peoples, O'Dwyer, Shields, & Wang, 2013), or to examine the predicative validity of math achievement (Vista, 2016).

DIF analysis. Another extensive usage of the Rasch family model is to detect differential item functioning (DIF). This involves investigation of the invariance property of the instrument. If an item or the test provides advantage to a particular subgroup of the examinees over another subgroup, it may introduce bias and impact fairness of the measure. For instance, gender difference on math ability was inspected through the Rasch lens (Liu & Wilson, 2009). Gender differences in favor of males were found on complex multiple-choice items as well as Space and Shape items in the U.S. portion of the Programme for International Student Assessment (PISA; OECD, 2000).

Other uses of the Rasch model are varied. These include investigations of math test scores across 21 nations (Yildirim, Yildirim, & Verhelst, 2014) and how explanatory models based on Rasch theory may be used to detect systematic differences between subgroups of test takers as well as subsets of items (Engelhard, Wind, Kobrin, & Chajewski, 2013). Rasch may be used to search for patterns in model-data fit for

differences in subgroups of test takers and items (Engelhard, Kobrin, & Wind, 2014), possible effects of booklet design (Harsch & Rupp, 2011; Hecht, Weirich, Siegle, & Frey, 2015), and effects of item position on student performance (Hohensinn, Kubinger, Reif, Schleicher, & Khorramdel, 2011). The use of Rasch model was also extended to the longitudinal investigation of educational outcomes for difference in group-specific growth (von Davier & Xu, 2011).

Reliability improvement and bias review. The Rasch model has been used in large-scale assessments to improve reliability of test scores and reduce bias for test takers. For example, the Rasch many-faceted model was employed to calibrate raters and tasks (Engelhard, 1992), to validate two consecutive administrations of an assessment (Bonk & Ockey, 2003), to cross-validate and calibrate an instrument with numerous moderator variables and ability dimensions (Kubinger et al., 2011), to calibrate scores from different raters in order to improve rater reliability (Congdon & McQueen, 2000; McQueen & Congdon, 1997), and to detect common patterns of rater errors (Wolfe & Chiu, 1997). Furthermore, researchers have used the Rasch model to examine guessing bias (Andrich, Marais, & Humphry, 2016) and to examine how eliminating guessing bias affects difficulty parameters (Andrich & Marais, 2018)

Miscellaneous usage. Studies have employed the Rasch family models for various other purposes such as to equate scores (Bowe & Cronin, 2005), to obtain examinee ability and test item estimates (Fleckenstein, Leucht, Pant, & Köller, 2016), to differential students on their attained level of competences (Mesic & Muratovic, 2011), or to analyze test results (Wolfe & Miller, 1997). Rasch models have also been used to analyze psychometric properties of instruments in pilot studies before the instrument is

fully implemented (Hagquist & Hellström, 2014), and to analyze factors in order to discover dimensionality of a latent construct (Sälzer & Heine, 2016).

Educational decision making. With the wide application of Rasch measurement models in various disciplines, Rasch analysis results have been used to improve assessment practice (Bonk & Ockey, 2003; Hecht et al., 2015; Kubinger et al., 2011; Liu & Wilson, 2009; Wolfe & Miller, 1997), to predict examinees' future performance (Bowe & Cronin, 2005; Fleckenstein et al., 2016), and to guide future instruction (Vista, 2016). Furthermore, educators have utilized Rasch study results to improve rater reliability through monitoring or training (Congdon & McQueen, 2000; Eckes, 2008; McQueen & Congdon, 1997; Wind & Engelhard, 2012), to improve result reporting (Engelhard et al., 2014), even to set cut score or benchmarking standards (Harsch & Rupp, 2011; Popp & Ryan, 2002; Wyse, 2011).

Rasch-based Fit Indices

As a popular tool that serves multiple measurement purposes, Rasch measurement models hold an important seat in large-scale assessments. As this method is commonly used in state-wide assessment, it is imperative to evaluate how this tool functions under different situations.

Use of fit indices. As a measurement model that is dedicated to the construction of fundamental measures (Bond & Fox, 2012), the Rasch model provides information on the performance of various model components to safeguard validity associated with the test scores. Fit indices (e.g., overall model fit, item fit, person fit) are used to evaluate these components, and these indices have been utilized for various purposes during the

test development process in different disciplines such as education, public health, psychology, and language arts.

Beginning in the 1970s and the early 80s, the work of Rasch-based fit index investigation was undertaken in the University of Chicago with researchers like Mead (1976) and R. Smith (1982). It was concluded that fit indices can be used for detecting measurement disturbances. For instance, the R. Smith simulation study (1982) compared the likelihood ratio fit statistic and the INFIT and OUTFIT between-fit statistic. Results indicated that the two Rasch-based fit statistic highly correlated with the likelihood ratio statistic (i.e., $r = 0.99$).

Using item fit statistics, researchers have identified items that did not appear to function satisfactorily to either improve the assessment (Baghi, 1990; Klassen et al., 2014; Papadopoulos, Spanoudis & Kendeou, 2009) or to reduce the number of items (Erhart et al., 2009; Kang et al., 2018). Pilatti and her colleagues used person and/or item fit information to examine the psychometric properties of the Spanish version of an alcohol consequences questionnaire (Pilatti et al., 2014; Pilatti, Read, & Caneto, 2016), an alcohol expectancies questionnaire (Pilatti, Godoy, Lozano, & Brussino, 2015), as well as an impulsive behavior scale (Pilatti, Lozano, & Cyders, 2015) during the scale development process.

Another major purpose of using Rasch-based fit indices is to provide validity support. Researchers and other stakeholders have used global fit indices (e.g., overall fit statistic) as well as other statistics (e.g., reliability) to validate a scale under investigation. For instance, overall fit, averaged person fit and item fit were analyzed to examine dimensionality and validate the construct of phonological abilities in Greek

(Papadopoulos et al., 2009). With an undergraduate admission test, the person fit statistic was examined to detect how individuals' profile conformed to the model to validate the assessment under repeated test administrations (Andrich, Styles, Mercer, & Puddey, 2017). In another study, item fit information was provided as validity evidence for a morale scale in a Hong Kong Chinese elderly population (Wong, Woo, Hui, & Ho, 2004). Additionally, global and item fit, along with age or gender DIF information, were studied to identify best items and validate a generic scale that measures self-management skills (Klassen et al., 2014).

As test/scale developers and Rasch researchers strive to create better measures with solid psychometric properties, fit indices aid them in achieving this goal. For better practice, educational researchers Walker, Jennings, and Engelhard (2018) investigated person misfit to improve score interpretation as well as detect potential threats to validity. All the above-mentioned studies have only reported a portion of how Rasch fit indices were utilized, but the research has provided ample examples on how these indices can be applied.

Specifically, each item fit statistic consists of three types: total item fit, between-item fit, and within-item fit, depending on how the squared standardized residuals are summarized (Smith, 2000).[1] Each type of item fit also has two versions, related to the use of weighting : weighted item fit and unweighted item fit. Furthermore, both the weighted and the unweighted versions of item fit can be calculated through mean square (MNSQ) values, or *t*-transformation standardized (ZSTD) values. Therefore, there are 12 variations of item fit statistics (e.g., weighted total item fit mean square value, weighted

[1] For detailed explanations and calculation methods, please consult Smith (2000).

total item fit standardized value, unweighted total item fit mean square value, unweighted total item fit standardized value, etc.). For this particular study, only the weighted and unweighted total item fit (MNSQ and ZSTD) as well as the unweighted between-item fit statistics (MNSQ and ZSTD) were used.

Item fit statistic. Within the Rasch framework, a fit statistic undertakes two primary branches of development: the Pearsonian chi-square approach (Wright & Panchapakesan, 1969) and the likelihood ratio chi-square approach (Andersen, 1973; Gustafsson, 1980; van den Wollenberg, 1982). The Pearsonian chi-square test evaluates the likelihood of obtaining the observed difference by chance; while the likelihood ratio chi-square test compares the goodness of fit of the null hypothesis model and an alternative hypothesis model. In Rasch measurement, the former is prevalent within the United States and Australia (Smith, R., 2004a).

The main item fit statistic, a Pearsonian chi-square value initially proposed by Wright and Panchapakesan (1969), has been used to identify items that do not fit the Rasch model (i.e., misfitting items). The chi-square statistic was based on the differences between the observed and the expected score for a group of test takers that have obtained the same raw test score (Smith et al., 1998). Following this concept, fit statistic was subsequently developed on the basis of the item/person residual, defined as the differences between the observed and the expected responses (Smith et al., 1998).

The fit statistic focuses on two versions: the information-weighted fit (INFIT) statistic and the unweighted fit (OUTFIT) statistic. The INFIT statistic is a weighted (WT) fit statistic, meaning that it gives more weight to the person with measures that are closer to the item difficulty value. Its mathematical formula is:

$$x^2\ (WT)_i\ =\ \frac{\sum_{n=1}^{N}(x_{ni}-p_{ni}\ )^2}{\sum_{n=1}^{N}(p_{ni}\ (1-p_{ni}\ ))}\,, \qquad [5]$$

where x_{ni} is the observed response of item *i* from examinee *n* and p_{ni} is the probability of examinee *n* correctly answering item *i* (Smith, R., 2004a, p. 83).

For the INFIT statistic, the weights (WT) are used to reduce the influence of responses that are less informative (i.e., low variance), or off-target (Wright & Masters, 1990). In the weighted formula above [5], information function ($p_{ni}\ (1-\ p_{ni}$)) is used to weigh each squared standardized residual value before summation. The paradox for the INFIT statistic is that regardless of the person and the item locations on the variable, any response is equally likely; this diminishes the power of the INFIT statistic to detect anomalous responses.

The OUTFIT statistic is unweighted (UT). Here, the formula simply sums all the chi-square values for any identified person-item interactions. As a result, the OUTFIT value is more sensitive to outlying person scores or unexpected observations (i.e., random disturbances) like guessing or carelessness (Smith et al., 2000). Therefore, OUTFIT problems are easier to detect and manage, and they create a lesser threat to measurement than the INFIT statistic (Linacre, 2002b). Following the mathematical formula by Smith et al. (1998), it can also be expressed as:

$$x^2\ (UT)_i\ =\ (\ \frac{1}{N}\ )\sum_{n=1}^{N}\frac{(x_{ni}-p_{ni}\ )^2}{p_{ni}\ (1-p_{ni}\ )}\,. \qquad [6]$$

Both INFIT and OUTFIT statistics can be expressed either in mean square (MNSQ) values or standardized values (ZSTD). Mean square values are calculated by taking the average value of the squared residuals for a particular item because the mere sum of all the residuals results in zero. The ZSTD values are calculated by dividing each

score residual by its standard deviation (SD) $(p_{ni}\ (1\ -\ p_{ni}))^{1/2}$ (Smith, R., 2004a). Mathematically, the formula for a standardized residual is:

$$y_{ni}\ =\ \frac{(x_{ni} - p_{ni}\ )}{(\ p_{ni}\ (1 - p_{ni}\ ))^{1/2}}\ , \qquad [7]$$

where x_{ni} is the observed response of item *i* from examinee *n* and p_{ni} is the probability of examinee *n* correctly answering item *i* (Smith, R., 2004a).

Symmetrical threshold values ignore the distributional properties of the MNSQ statistic, suggesting the MNSQ statistic is not symmetrical. The expected value of this statistic is 1.0, with a minimum value of 0.0 and a maximum value of + ∞. The shape of the distribution precludes the use of symmetrical critical values for their upper and lower tails of the distribution as demonstrated by the Smith et al. study (1998). Linacre (2019a) explicitly explained how to use these statistics to diagnose item fit. Larger MNSQ values indicate a larger difference between the expected responses to the item and actual responses. Values that are larger than 1.0 suggest underfit to the model meaning that there is unmodeled noise or other sources of variance and the data are less predictable than expected. Values less than 1.0 indicate overfitting and data are more predictable than expected. A mean square of 0.6 indicates the data are too "Guttman-like" (suggesting a predictable hierarchical order) and there is a 40% deficiency in the predicted randomness. This implies that there is 100 x (1 − 0.6) / 0.6 = 67% more ambiguity in the inferred measure (e.g., a scale or instrument to measure a latent construct) index than is modeled, suggesting "a few random responses from low performers", or mis-performing items by respondents who the items are aimed for (Linacre, 2019a). Standardized values are converted through the cube-root

transformation to approximate a normal distribution and are reported as a *t*-statistic with infinite degrees of freedom.

In general, MNSQ values between 0.7 and 1.3, as well as ZSTD values of −2.0 to +2.0 are considered acceptable for both indices (Wright & Linacre, 1994). Mean square values below 0.7 may produce excessively high reliability and separation coefficient (i.e., a correlation coefficient that classifies a group of examinees or verifies the hierarchy of test items; Linacre, 2019b) values that are misleading; whereas values larger than 1.3 may distort or degrade the measurement system, compromising unidimensionality of the instrument (Linacre, 2002b; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). Instead, more conservative MNSQ values of 0.8 and 1.2 are suggested under high-stakes situations (Wright & Linacre, 1994).

For ZSTD, standardized values of −3.0 to +3.0 have been recommended as cut-off values for large samples (Iramaneerat, Smith, E., & Smith, R., 2007). Standardized values below −2.0 suggest that data are too predictable and that other "dimensions" may be present to constrain the responses (to make the response patterns predictable). Values that are larger than +2.0 suggest that the data are noticeably unpredictable and, thus, do not provide much useful information. According to R. Smith (2004a), ZSTD values larger than ±2.0 will occur only when the probability of a correct response is less than 0.2 or greater than 0.8 and when the response is in the unexpected direction. For values greater than ±3.0 to occur, the probability of a correct response is less than 0.1 or larger than 0.9 and the response is in the unexpected direction.

While INFIT and OUTFIT can be computed for both persons and items, here, the focus is on item fit because in general researchers are more concerned with item fit for

large-scale testing (and subsequent validation) purposes. As there are 12 different fit indices, the indices to be used in this study are defined. The focus of Research Questions 1 and 2 are to investigate the total item weighted (INFIT) and total item unweighted (OUTFIT) fit statistics, both MNSQ and ZSTD values. For Research Question 3, the unweighted (OUTFIT) between-item fit MNSQ and ZSTD values were used for the power analysis due to their sensitivity to systematic measurement disturbances and availability in Winsteps® procedures.

As noted earlier, when researchers use total item fit statistics to detect problematic responses to items, both the MNSQ values and the ZSTD values present a dilemma. Mean square values take into account the sample size and these values will approach 1.0 (illustrating good model-data fit) simply by increasing the sample size. When the sample size is large enough, most (if not all) items would 'appear to' fit well. If ZSTD values are used instead, even the smallest amount of misfit becomes significant (i.e., > 2.0) when the sample size is large enough. Therefore, most items will likely be rejected based solely on the values of fit indices (Bond & Fox, 2012).

Based on this conclusion, in large-scale testing situations whereas the number of examinees easily exceeds a few thousand, it is very likely that most items would appear to fit well if MNSQ values are used, or most items would likely be rejected if guidelines of ZSTD values are followed. Although ZSTD may perform better than the MNSQ when large samples (i.e., approximating empirical testing situations) are used, there has been limited research involving both under such situations. Due to the relative lack of computation ability of computers and calibration programs to deal with large sample sizes with much of the Rasch fit investigations conducted in the 1980s, so far, only one

study on total item fit indices (DeMars, 2017) was conducted with a sample size of 5,000. All the other related research has investigated sample sizes no larger than 2,000 test takers; however, large-scale testing programs typically include tens of thousands of test-takers.

Impact of measurement disturbances. Fit statistics are used to detect high-quality measurement items that provide accurate information on student performance. Measurement disturbances introduce construct-irrelevant variance in student performance and may distort inferences drawn from assessment results. A frequently occurring type of measurement disturbance creates systematic errors and favors a particular subgroup of examinees after their ability levels are matched. Items that provide advantage to certain examinees due to their group membership are deemed biased, thus it is imperative that the fit statistics function well in such situations to detect the items that function differentially, in other words, Differential Item Functioning (DIF). DIF analysis is usually conducted to evaluate the invariance property of a measure across subgroups of test takers (e.g., racial/ethnic or gender subgroups). Various methods (e.g., Mantel-Haenszel [MH] statistic, Mantel & Haenszel, 1959; SIBTEST, Shealy & Stout, 1993; item-focused tree permutation test, Tutz & Berger, 2016) have been developed to detect DIF. Under the Rasch model, the between-fit statistic is generally recommended to evaluate if an item functions differently between subgroups of examinees (Smith, R., 2004a), and the rule-of-thumb guidelines (i.e., 0.7 and 1.3 for MNSQ and ±2.0 for ZSTD in non-high-stakes situations as suggested by Wright & Linacre, 1994) are usually followed for item evaluation.

Differential Item Functioning has been well studied, especially in large-scale testing programs. A meta-analysis study on reading assessment (Koo, Becker, & Kim, 2014) concluded that English Language Learning (ELL) students and non-ELL students performed differently on different types of reading comprehension questions in third and tenth graders when their gender and ethnic backgrounds were taken into consideration. Further, using test results from a state-wide mathematics assessment, researchers (Scarpati, Wells, Lewis, & Jirka, 2011) noted that performance differences were associated with item difficulty, student ability, as well as accommodation provided to students with disabilities. A study on the minority and majority of Francophone students in Canada suggested the existence of heterogeneity of DIF within each linguistic group; and researchers proposed developing "tests that are fair for a subset of examinees" instead of all subgroups due to the detected within-group heterogeneity (Oliveri, Ercikan, Lyons-Thomas, & Holtzman, 2016). Cross-nationally, German researchers (Sachse, Roppelt, & Haag, 2016) identified the presence of DIF (e.g., item translation, cultural specificities, curriculum coverage) in the Programme for International Student Assessment (PISA; OECD, 2000).

Given the importance of item quality to test validity, it is important to understand how the item fit statistics function in detecting misfitting items. In other words, it is also imperative to test the ability (or the power) of the suggested fit statistic used to identify measurement disturbances. It is hoped that the fit statistic obtained from the proposed study would perform well in identifying the items that present DIF, which may be introduced by gender, ethnicity, social economic background, etc. Ideally, the results

would reach a satisfactory accuracy rate indicated by the percentage of true positive cases.

Research on Item Fit Indices

The general guidelines on fit statistics proposed by Wright and Linacre (1994) have been popular among psychometricians and researchers. However, in the field of education, only a handful of studies have been conducted to validate the suggested values. A simulation study by Smith and colleagues was conducted in the late 1990s and only a couple of studies focused on this topic in the 2000s (Wang & Chen, 2005; Su, Sheu, & Wang, 2007). This topic has recently gained attention, and resulted in several additional studies (DeMars, 2017; Hodge & Morgan, 2017; Seol, 2016; Wolfe, 2013).

Using BICAL and BIGSTEPS, early Rasch researchers Smith and colleagues (1998) studied the MNSQ values as well as the ZSTD values. They concluded that the former appeared to be more sensitive to factors such as sample size than the latter; therefore, suggested ZSTD values be considered to identify item misfit. Furthermore, different critical values for total fit indices may need to be considered as a function of sample size. For instance, to have a consistent Type I error rate of approximately .05, unweighted mean square critical values of 1.3, 1.2, and 1.1 would be needed to identify misfit for sample sizes of 150, 500, and 1,000, respectively (Smith et al., 1998). This study suggested that using a single critical value for the mean square to identify item misfit can lead to under- or over- detection (Smith et al., 1998). Also, results from this study showed that the MNSQ values obtained by the study are not symmetrically distributed, and researchers suggested that symmetrical critical values for detecting item

misfit would have different Type I error rates for the upper and lower tails of the distribution.

In 2005, Taiwanese researchers Wang and Chen conducted a study using Monte Carlo simulations. In the Rasch dichotomous model, they manipulated eight levels of sample sizes (100, 200, 400, 600, 800, 1000, 1,500, and 2,000 examinees) and four levels of test lengths (10, 20, 40, and 60 items) to investigate how item parameter recovered, as well as the performance of standard error estimates and total fit statistics. Results showed that for the item parameters, Winsteps® yielded biased estimates. In short tests (20 items and fewer), Winsteps® overspread the difficulty estimates: it underestimated easy items yet overestimated difficult items. However, the parameter estimates tended to be non-substantial in longer tests. After the researchers implemented a correction procedure introduced by Wright and Douglas (1977) through multiplying the correction factor of $(L - 1) / L$, whereas L being the test length, the magnitudes of the estimation bias were considerably reduced to close to zero[2] (Linacre, n.d.).

Although Winsteps® did not yield biased standard error estimates, it yielded INFIT and OUTFIT mean square values that were approximately equal to 1.0 on average. Even though the standard deviations of these values failed to converge to a constant, they tended to decrease as the sample size increased. Furthermore, the standard deviations of the OUTFIT mean square values were larger compared to those of the INFIT values. While the sample size appeared to impact the estimates, the test length did not seem to affect both the item INFIT and the item OUTFIT mean square values. This study also suggested that both the INFIT and OUTFIT mean square values showed different

[2] Winsteps® corrects this bias with the normal statistical correction to include estimates of the item difficulty.

magnitudes of variability as sample sizes changed, and they needed different critical ranges to screen misfitting items.

This simulation study also showed that the means of the INFIT and OUTFIT standardized values were very close to the theoretical value of zero, yet their standard deviations were smaller than their theoretical values of 1.0 (Wang & Chen, 2005). Again, the standard deviations of the OUTFIT standardized values were larger than their counterparts of the INFIT values. The researchers suggested that in general, the commonly used critical ranges of ±2.0 at the .05 level were safe to use to screen items with moderate difficulties and smaller sample sizes ($n < 2000$). However, this range would be too conservative when items were extremely easy/difficult. Using this range might result in incorrectly identify poor-fitting items. Therefore, the researchers proposed the use of a correction so that the INFIT and OUTFIT standardized values would more closely follow the standard normal distribution. Results from Wang and Chen (2005) indicated that the standardized values should be considered to detect item misfit, as suggested by Smith (1982).

Another study conducted by Su et al. (2007) used the parametric bootstrapping method to examine a set of 15 items with two levels of sample sizes of 200 and 2,000. Their results suggested that point estimates of the MNSQ values alone were not appropriate for use to identify misfitting items. Furthermore, researchers concluded that it was inappropriate to use the rule-of-thumb cut-off values of the standard normal distribution to screen items because the INFIT and OUTFIT standardized values do not exactly follow the standard normal distribution. Similarly, researchers reiterated that sample size influences the use of INFIT and OUTFIT MNSQ values. When a sample

size became sufficiently large, all items in a real test would be eventually identified as poorly fitting. Therefore, they suggested, we needed to use effect-size measures to assess the magnitude of the misfit.

Fast forwarding to the 2010s, Wolfe (2013) also employed the bootstrap procedure to evaluate the total fit statistics to the family of Rasch models. For the simulated dichotomous model, he generated data from a standard normal distribution [i.e., N (0, 1)] for the 1,000 trait scores within the < 2,000 range for ±2.0 critical values and from a uniform distribution [i.e., U (−2, +2 logits)] for the 100 different values of item difficulty. All data were randomly sampled from the associated distributions.

Item fit results showed that the bootstrap critical values performed quite well with respect to Type I error rates for the INFIT mean square and standardized values. Values were slightly lower for the two versions of the OUTFIT index, but were still fairly close to expected values.

Using empirical data and a dichotomous model, researchers resampled from 385 respondents answering a 58-item self-reported questionnaire (Viger, Wolfe, Dozier, & Machtmes, 2006). Results suggested that the bootstrap critical values for the INFIT mean square and standardized values flagged more items than the widely used values; whereas the OUTFIT mean square and standardized values identified fewer items than the often-used values. This study reiterated that the ZSTD values of both the INFIT and the OUFTIF total item fit statistics tended to perform better than their MNSQ counterparts.

As researchers asserted that point estimates may not be appropriate to detect ill-fitting items, some studies (Su et al., 2007; Wolfe, 2013) proposed expanding the

estimates to a range of values, i.e., using confidence intervals (CIs), an interval or the boundary of numbers the possible values of this statistic are believed to fall within (Agresti & Finlay, 2009). Using the bootstrap approach, Seol (2016) investigated the pattern of the bootstrapped CIs due to sample sizes and test lengths with polytomous data. A total of 25 testing conditions resulting from five levels of sample size (from 200 to 1,000 by an increment of 200 cases) and test length (10, 20, 40, 60, and 80 items) were examined. Study results indicated that mean and standard error values were fairly close to their theoretical values of unity and zero for MNSQ statistic, INFIT, and OUTFIT indices. The study also suggested that the 95% CIs for the INFIT and OUTFIT mean square values became narrower as the sample size and the test length increased. The CI values for the standardized version of these statistics were fairly stable across different sample size and test length conditions. Even though variability of the INFIT and OUTFIT mean square values were partially consistent with the Wang and Chen's study (2005), the bootstrapped critical values were still different from the conventional criterion of (0.6, 1.4) for a polytomous scale suggested by Wright and Linacre (1994). The researcher also concluded that the bootstrap critical values could be used to detect misfitting items because they offered a reasonable alternate.

To investigate the magnitude of item misfit, DeMars (2017) utilized the conjunctive tests of statistical significance and effect size. Her simulation study contained 46 items and two levels of sample size: 100 and 5,000. The majority (i.e., 40) of the 46 simulated items fit the Rasch model, and six items were misfitting with various issues, e.g., additional item parameter, extremely easy, off-target to the ability estimates, etc. Results showed that the means and medians of the INFIT mean square values

(MNSQ) were the same across sample sizes for well-fitting and poorly-fitting items. Even though the OUTFIT mean square values yielded the same mean and median values for items that fit, these values appeared to be influenced by the sample size in the worst-fitting items. The Type I error rates, based on the ZSTD values, were used to flag items based on statistical significance instead of magnitude of misfit. Results suggested that the ZSTD values of the INFIT and OUTFIT statistics reasonably followed a standard normal distribution in middle difficult items. The extreme items resulted in the departure from the normal distribution in the tails, and the rejection rates were somewhat lower than the traditional 2.5% for a two-tail distribution. When both the MNSQ values and the ZSTD values were taken into account, the researcher concluded that the MNSQ indices could be used as effect size measures along with statistical significance information from the ZSTD indices for a better hit rate.

The most recent study on fit statistic in educational research was conducted by Hodge and Morgan (2017). This study investigated the stability of the INFIT and OUTFIT MNSQ estimates as compared to the commonly used rule-of-thumb critical values. This simulation study used the applied data set from a clinical exam for the American Chiropractic Neurology Board Certification which involved 196 candidate examinees and 215 dichotomous test items. Results from this study were consistent with some previous studies and the researchers concluded that the rule-of-thumb critical values suggested for fit statistics may be inappropriate in detecting item misfit.

As the use of Rasch models has extended from educational assessment to health research, medical researchers drew samples with replacement from two Likert scale instruments and investigated the relationship between total fit statistics and sample size

for two polytomous models: the rating scale and the partial credit (Smith, A. B. et al., 2008). With the eight sample sizes from 25 to 3,200 with the latter sample size being twice as much as the previous one (i.e., 25, 50, 100, etc.), the study concluded that MNSQ values were relatively invariant of sample size for polytomous data, whereas ZSTD values were highly sensitive to sample size. It also suggested that mean square statistic could identify item misfit to the model using published recommended ranges of (0.7, 1.3).

In summary, the previous research on Rasch total item fit statistics appears to endorse the use of ZSTD values over the MNSQ values to identify misfitting items; however, researchers still cautioned that the traditional cut-off values of ±2.0 for ZSTD might not always be appropriate for screening (Wang & Chen, 2005). Results from the collection of studies suggest that the ZSTD values of the total item fit statistic appear to be more stable in comparison to the MNSQ values. In addition, the studies found that the MNSQ values of both INFIT and OUTFIT statistic yielded variability that changed as the sample size increased (DeMars, 2017; Hodge & Morgan, 2017; Seol, 2016; Smith, R. M. et al., 1998; Su et al., 2007; Wang & Chen, 2005) and test length increased (Seol, 2016). Meanwhile, the *t*-transformation ZSTD values remained more stable across different sample size and test length conditions but did not follow a normal distribution (Seol, 2016; Wang & Chen, 2005). However, a different conclusion was observed where MNSQ statistics were favored over their ZSTD counterparts in polytomous data because the latter were highly sensitive to sample size (Smith, A. B. et al., 2008).

While these studies have informed the field, there are shortcomings that are apparent. In these studies, most sample size of examinees ranged from 100 to 3,200 and

majority of the test length conditions focused on fewer than 60 items. However, the number of students who participate in a state-level testing is typically much larger, ranging from 1,000 to over 50,000. Additionally, state standardized testing programs typically employ multiple-choice tests that range between 45 items (e.g., 4th grade Science in SCPASS in 2015) to 66 items (e.g., 8th grade Social Studies in SCPASS in 2015). Thus, investigation of larger sample sizes and longer tests is needed to provide accurate recommendations to the field. Additionally, none of these previous studies has further investigated how well the suggested statistics perform in screening items with misfit (i.e., detecting measurement disturbances).

Purpose of the Present Study

The current study proposes to further investigate how Rasch item fit statistics (i.e., the total item INFIT and OUTFIT mean square values [MNSQ], the total item INFIT and OUTFIT standardized values [ZSTD]) performed, under situations that mirror those found with large-scale assessment settings. Findings from previous studies suggest that the widely used rule-of-thumb critical ranges are not appropriate to detect misfitting items, and the literature has remained inconclusive what cut-off values should be used to detect misfitting items. Further, it is still unknown what the associated Type I error rates are for the total item fit mean square values. In addition, the question of accuracy of all these values in detecting measurement disturbances remains. The proposed study attempts to address these three questions in an hope to shed new light on the use of these Rasch-based item fit statistics. Additionally, as systematic measurement disturbances were simulated (i.e., DIF investigation), between-item fit statistics OUTFIT MNSQ and

ZSTD were used to investigate the power and error rates associated with the obtained critical values.

When the stakes are high for state-wide assessment programs, it is imperative that items with solid psychometric properties are selected and included in the assessment. As this study investigates item fit statistic with situations that are typically encountered in large-scale testing situations, the results can provide valuable information regarding if item performance is as anticipated.

CHAPTER 3

METHODOLOGY

Item quality is vital in developing an efficient and fair test that provides stakeholders consistent and useful information about test takers. When the Rasch model is used, traditionally, the mean square (MNSQ) values and the standardized (ZSTD) values of the weighted and the unweighted total fit indices have been used to evaluate item quality. However, studies have found some results from simulated data on accuracy of these indices and questioned the general rule-of-thumb cut-off guidelines suggested by Wright and Linacre (1994). The purpose of the current study was to investigate which fit indices were more accurate in detecting misfitting test items in large-scale testing, how accurate they were, and how well they could detect measurement disturbances.

This chapter will briefly describe the design of the simulation study and how data were generated. It will then detail why and how different levels in each condition were determined (e.g., number of items, number of test takers, etc.). Finally, the chapter will provide a brief analysis plan on how to evaluate different total item fit statistics and how to examine the power of the obtained values in detecting simulated systematic item misfit through between-item fit statistics.

Simulation Study

A simulation study was conducted to examine the impact of sample size and test length on total item fit statistics, when the Rasch dichotomous model was used for analysis. Both the MNSQ values and the ZSTD values of the INFIT weighted and the

OUTFIT unweighted statistics were investigated, as they are the widely used indices to detect misfitting items (Wolfe, 2013). To investigate total item fit indices, the WinGen program (v. 3; Han, 2007) was used to generate the initial response patterns for the original data set, and the Winsteps® software package (v. 4.4.8; Linacre, 2019c) was used for subsequent data simulation due to the additional data manipulation process from WinGen to Winsteps®.

To examine power of the obtained indices, R (v. 3.5.3; R Development Core Team, 2019) was applied to generate responses to items with and without measurement disturbances[3]. Winsteps® was then used in all item fit analyses including fit indices and power investigation. Finally, since Winsteps® only produces item fit statistical results on one particular data set, the SAS® software package (v. 9.4; SAS Institute Inc., 2013) was utilized to organize results from all replications as well as to calculate summary of results. In particular, a Macro (Miller, 2004) was included as part of the codes to summarize Winsteps® results.

The data sets constructed for the study used pre-specified parameters for person ability and item difficulty estimates. Specifically, person ability estimates were generated from a normal distribution with a mean of zero and a standard deviation of 1 [i.e., N (0, 1)], which corresponds to a commonly used population distribution. The item parameters were generated from a uniform distribution [i.e., U (−2, +2 logits)], which suggests items would fit the Rasch dichotomous model (Wright & Douglas, 1977).

Sample size and test length were varied to examine their impact upon the accuracy of different total item fit statistics, as well as their associated Type I error rates.

[3] R codes from a previous study (Pompey, Jiang, Burgess, & Lewis, 2019) on DIF analyses were modified for data generation.

The goal was to examine critical values and their frequencies of occurrence under various situations manipulated in the study. For the simulation, the critical values are closely related to Type I error rates and power when data are misfitting. Furthermore, measurement disturbances were simulated to investigate the ability of the chosen statistics in detecting these disturbances. The disturbances were simulated to approximate uniform DIF (i.e., the same DIF across all ability levels) and they were simulated in two-group differences with a balanced design (i.e., the same number of test-taking individuals in each group). To demonstrate magnitude of DIF, items exhibiting moderate to large DIF as categorized "C" in the Educational Testing Service (ETS) guidelines (Zieky, 1993) were set at varied percentages of items present in the test.

The ETS DIF analysis procedures use the Mantel-Haenszel (MH) statistic (Mantel & Haenszel, 1959) and a delta scale of item difficulty (Holland & Thayer, 1985) to categorize the magnitude of item DIF. Developed by Holland and Thayer (1985), the *MH D-DIF* index is calculate as:

$$MH\ D\text{-}DIF\ =\ -2.35\ln(\hat{\alpha}_{MH})\ , \qquad [8]$$

where the constant odds-ratio is estimated as (Mantel-Haenszel, 1959):

$$\hat{\alpha}_{MH}\ =\ \frac{\sum_K N_{R1K}N_{F0K}/N_K}{\sum_K N_{R0K}N_{F1K}/N_K} \qquad [9]$$

Note: N_{R1K} and N_{F1K} refer to the numbers of examinees in the reference and focal groups who answer correctly; N_{R0K} and N_{F0K} represent the numbers of examinees in the reference and focal groups who answer incorrectly; N_K is the total number of examinees.

A positive value of *MH D-DIF* suggests an item is estimated to be more difficult for the reference group (i.e., the group that is anticipated to be advantaged by a measure; "Differential Item Functioning", n.d.), whereas a negative value suggests an item is more

difficult for the focal group (i.e., the group that is suspected to have a disadvantage in a measure; "Differential Item Functioning", n.d.) and favors the reference group (Zwick, 2012).

Based on the magnitude of this index as well as its statistical significance, items with this statistic not significant at the .05 level or its absolute value is smaller than 1 are categorized as "A" (negligible or nonsignificant DIF). Items with the *MH D-DIF* significantly greater than 1 in absolute value at the .05 level as well as an absolute value of 1.5 or greater are considered "C" (moderate to large DIF). Items that do not meet the criteria for "A" or "C" are deemed having slight to moderate DIF (Zwick, 2012). The ETS guidelines recommend test developers not to select any "C" items unless they are essential for important test specifications and the factors contributing to their high DIF level are determined not to represent bias (Zieky, 2003). Category "C" items are usually revised or excluded from future test administrations; therefore, it is imperative that obtained statistics can accurately capture these items.

Data Generation

Parametric bootstrapping was used to generate data. Introduced by Efron (1979), the bootstrap method is a statistical procedure that relies on random resampling to obtain various estimates like standard errors, CIs, etc. As researchers (Hesterberg, Moore, Monaghan, Clipson, & Epstein, 2005) pointed out, distributions and quantities from bootstrapping do not differ much if a large number of resamples (typically about 1,000) are used to generate a probability distribution of a statistic of interest. The parametric bootstrap resamples from a hypothetical distribution and this method is known to "produce sampling distribution estimates that exhibit bias, spread, and shape similar to

that of the parametric sampling distribution (Wolfe, 2013, p. 4)". This method requires fewer assumptions to be met in estimating sampling distributions (e.g., distributions do not need to be normal, sample sizes from empirical data do not need to be large; Hesterberg et al., 2005).

The bootstrapping method allows researchers to derive estimates for complex estimators of complex parameters of the distribution (Hilmer & Holt, 2000). A major advantage of this method is the appropriateness to control and check the stability of the results. They are asymptotically consistent, suggesting that the estimates of the asymptotic distribution of a parameter are consistent (Ritesh, 2016). While it is impossible to repeatedly sample a large number of examinees to take and retake an assessment, simulating the data in various testing situations renders researchers a realistic tool to calculate the statistics of interest. Furthermore, employing bootstrapping methods to randomly sample replacement data can provide consistent results.

Nevertheless, the bootstrapping procedure used in Winsteps® creates the bounds of the population given the target characteristics and draws unique samples of a specific size from the pre-determined distribution in the data generation. The stochastic structure of the data is necessary for the Rasch-based fit statistics to function accurately. In other words, sampling needs to be a random process for unique samples to be representative of the larger population. Violation of this process will result in over- or under-sampling of certain classes/categories of a data set, thus impact the accuracy of data analysis. For instance, if a sample of size 100 was drawn and then multiplied by 500 in order to obtain a sample of 50,000, this will limit the performance of fit indices as the data generated is limited in terms of variability, response conditions, and person/item characteristics. The

end result of such a procedure would adversely affect accuracy of the results (R. Smith, personal communication, July 15, 2020). As the Users' Guide states, the procedure used by Winsteps® creates resamples "with replacement from the current dataset" (Linacre, 2019d), and this might impact the accuracy of results. Details of the simulation study follow.

Tested fit statistics. Generally, the weighted total item fit statistic (i.e., the INFIT) is more sensitive to unexpected response patterns on items when items are roughly targeted on the test taker's ability level; and the unweighted total item fit statistic (i.e., the OUTFIT) is more sensitive to unexpected responses on items that are extreme compared to the test taker's ability as being too easy or too difficult (Linacre, 2019a). These fit statistics used are popular indices favored by psychometricians and researchers to identify misfitting items under the Rasch model.

Historically, the established rule-of-thumb values for the MNSQ fit statistics consider 0.7 for the lower limit and 1.3 for the upper limit. Items with values beyond these limits are deemed misfitting to the model. Under high-stakes situations, these values may be altered to bounds of 0.8 and 1.2 (Wright & Linacre, 1994). Additionally, the traditional ±2.0 are adopted as the limits for standardized values to detect misfit (Wright & Linacre, 1994). For large sample sizes, the values of ±3.0 are recommended (Iramaneerat et al., 2007). However, previous simulation studies investigating the total fit statistic (Hodge & Grant, 2017; Wolfe, 2013) have concluded that the rule-of-thumb guidelines were inappropriate in many applied situations. Using the MNSQ values may result in under-identifying misfitting items while the ZSTD values might over-identify misfitting items when the sample size is large (Bond & Fox, 2012). Furthermore, the

suggested MNSQ values indicate that their distribution is not symmetrical; therefore, different critical values are needed for the upper and the lower ends (Smith et al., 1998). The current study attempted to simulate testing situations with various large numbers of test takers and multiple-choice items to investigate how the suggested total item fit indices function. Also, their associated critical values were examined to obtain the associated Type I error rates for these testing situations. Moreover, measurement disturbances (e.g., a two-group DIF like in gender) were simulated to test the ability of the obtained critical values to detect the items demonstrating differential functioning across groups.

Test length. The currently adopted assessments in South Carolina use multiple-choice items in most state assessments and all the national assessments. In test design, tests of 30 to 60 multiple-choice items appear to be commonly used for state tests (Burton, 2006). To mirror empirical testing situations, the total number of items for the original data of this study included three conditions: 30, 50, and 70 multiple-choice items, which suggests medium to long tests. The items were randomly sampled from a uniform distribution that ranges from −2 and +2 of the logit scales.

Sample size. In this study, subject sample size is important because chi-squared related test statistics used to evaluate item quality are affected by the number of test takers. According to the 2012-2013 National Center for Education Statistics (NCES, 2017), the number of public high school graduates ranged from 3,961 (District of Columbia) to 422,125 (California). Thus, in a typical state-wide assessment situation, the number of test takers can reach from a few thousand to hundred-thousands of students. For instance, South Carolina students at the elementary levels (3rd to 5th grades) take the

College-and-Career-Ready Assessments (SC READY; SC Department of Education, 2016), Palmetto Assessment of State Standards (SCPASS; SC Department of Education, 2008), samples of them are selected to take the National Assessment of Educational Progress (NAEP; National Center for Education Statistics, 1969), and some may take Measures of Academic Progress-Reading and Mathematics (MAP®; Northwest Evaluation Association, 2013). These students may participate in other district specific testing programs. In addition to some of these, some advanced middle level students (6th to 8th grades) take End-of-Course Examination Program (EOCEP; SC Department of Education, 2006), and high school students (9th to 12th grades) take the Scholastic Assessment Test (SAT; College Board, 1993) or the American College Testing (ACT®; ACT Inc, 1959) for college entrance qualification, in addition to EOCEP if not taken in middle school.

A quick review of the SCPASS program during the time period from 2009-2015 showed that the number of test-taking students were over 50,000 for Writing, English Language Arts (ELA), and Math in each of the 4-8 grades, over 20,000 for Science and Social Studies for Grades 3, 5, 6, and 8, and over 50,000 for Grades 4 and 7 in these two subjects. The number of test takers of the EOCEP and the High School Assessment Program (ELA and Math, HSAP; SC Department of Education, 2009) were also in the over 50,000 range from 50,747 (2011-12 HSAP Math) to 62,655 (2016-17 EOCEP Math). Review of testing situations within South Carolina helped to inform the sample size conditions used in this study. Technology has rendered an assessment team the ability to obtain test and item latency (i.e., time required to complete a test or answer a question item; Quan, Park, Sandahl, & Wolfe, 1984) information (Olsen, Maynes,

Slawson, & Ho, 1989), and a number of 40,000 to 50,000 test takers has become commonplace for item calibration/test equating (R. Smith, personal communication, December 12, 2019). Here, sample sizes of 5,000, 10,000, 25,000 and 50,000 were crossed for the three test length conditions.

Condition of measurement disturbances and DIF magnitude. There are various types of disturbances, some are random (e.g., guessing, carelessness) while others are more systematic (e.g., DIF, rater effect, test format). Even though systematic disturbances do not result in "inconsistent measurement", they may still "cause test scores to be inaccurate and thus reduce their practical utility" (Crocker & Algina, 2008, p. 106). DIF, one type of systematic error, also introduces construct-irrelevant variables that would impact accuracy of the statistical conclusion and score interpretation. The present study attempted to use simulated DIF measurement disturbances to evaluate how the obtained results perform in detecting poor-fitting items.

Various components can be designed in the item DIF, for instance, how DIF is distributed in different groups, how many subgroups with DIF are present in a given testing situation, how many test-taking subjects are in each group, etc. For this particular study, the item DIF were considered as uniform, suggesting "the statistical relationship between item response and group is constant for all levels of ability" (Awuor, 2008, p. 5). Uniform DIF indicates no interaction between group memebership and ability level in item response patterns, and when a DIF item favors one particular subgroup, only the item difficulty estimate (i.e., item *b*-parameter) differs among subgroups (Camilli & Shepard, 1994) as in Rasch models. When an item favors one particular subgroup, test takers from that group would have a better chance of providing a correct answer. In other

words, an item would be easier for a subgroup of test taker because the item is biased against test takers from the other subgroup (for instance, a reading question about car mechanism may favor male test takers vs female test takers). Also, the study used a balanced two-group design: one reference, and the other focal; and each subgroup had the same number of test-taking students because the presence of unequal sample sizes may requre statistical adjustment (Rusticus & Lovato, 2014).

Following Shepard et al.'s suggestion (Shepard, Camilli, & Williams, 1985), DIF was introduced through altering the item *b*-parameter (i.e., the difficulty estimate). The items that exhibit DIF were pre-selected and their percentage accounted for 4%, 10%, 20%, and 40% respectively of the total number of items. These percentages reflect the number of biased items "typically found in studies of ethnic, race, or gender bias (5% to 10%)" (Miller & Oshima, 1992, p. 383) as well as studies of instructional effects (20% to 40%) (Miller & Linn, 1988).

Another major component in DIF is its magnitude. Following the ETS' classification (Zieky, 2003), a "C" item exhibits moderate to large DIF. The group difference is expressed on the delta scale of item difficulty (i.e., MH D-DIF), and the "C" items have an MH D-DIF statistic that is significatenly greater than 1.0 and its absulte value of 1.5 or greater. Selected items exhibited a range of moderate to large DIF including 0.35, 0.45, 0.55 and 0.65 logits higher in item difficulty for the reference group. In other words, specific items were more difficult for the reference group. Overall, this translates into a total of 4 (sample size) x 3 (test length) x 4 (proportion of DIF items) x 4 (magnitude of DIF) = 192 conditions for DIF detection. Here, the between-group DIF was investigated to determine if items differ by subgroups.

Data Analysis

To assist in determining the item misfit for test development, this study examined the accuracy of total item fit statistics (i.e., MNSQ values and ZSTD values of INFIT and OUTFIT indices), their associated Type I error rates for the rule-of-thumb MNSQ and ZSTD cut-off values, and power of the obtained fit statics to detect simulated measurement disturbances. To examine the accuracy of total item fit statistics, first, the average fit statistic estimates across the test (i.e., the mean, standard deviation, the minimum value, and the maximum value) were computed per each condition. Then, these four descriptive indices from the 1,000 replications were calculated and examined across all test length and sample size conditions. These values provided an estimate of the variability across the estimates and were compared to the expected values for each fit index (i.e., mean value of 1 for MNSQ, and mean of 0 with standard deviation of 1 for ZSTD values).

The current study also examined the Type I error rates associated with the rule-of-thumb values suggested for assessment situations. This is notable, especially for the MNSQ values as the distribution of values is not symmetrical. If the error rates differ from the rule-of-thumb Type I error rate of .05, critical values that falsely identify item misfit at the 5% were examined.

To examine power, this study used the hit rate to assess the accuracy of identifying poorly fitting items. The hit rate was reflected by the true positive cases and it was calculated by summing the total number of correctly identified DIF items, dividing by the number of items simulated with DIF, and then averaging over the 1,000 simulated replications. The hit rate were calculated for each of the 192 (12 crossed sample size and

test length conditions x 4 percentages of DIF items in test x 4 magnitudes of DIF items) testing conditions, and a hit rate that reaches the 70% threshold was considered acceptable power for DIF detection as suggested by González-Romá, Hernández, and Gómez-Benito (2006). It was hoped that the obtained critical values would have adequate power to accurately detect poorly fitting items. Furthermore, the false positive rates (i.e., percentages of mis-identification of well-fitting items, Type I error rates) were calculated through summing the total number of incorrectly identified non-DIF items, dividing by the number of items simulated without DIF, and then averaging over the replications.

For indices to detect item DIF, the between-item fit item statistics for MNSQ and ZSTD were used instead of the total fit statistic as in the investigation of total item fit statistic accuracy and Type I error rates (i.e., Research Questions 1 and 2). The total fit statistic uses response information from all total scores while the between-item fit statistic takes into account differences in subpopulation membership (Smith, 1994). This is due to results from Smith (1994) which noted that the between-fit statistics were more sensitive to systematic measurement disturbances. Mathematically, the unweighted between-item fit statistic is expressed as:

$$x^2\ (UB)_i\ =\ \frac{1}{J-1}\ \ \sum_{j=1}^{J}\frac{(\sum_{n\epsilon j}^{N_j} x_{ni}\ -\sum_{n\epsilon j}^{N_j} p_{ni})^2}{\sum_{n\epsilon j}^{N_j}(p_{ni}\,(1-p_{ni}))}, \qquad [10]$$

where J represents the number of subpopulations, and N_j is the number of examinees in each subpopulation (Smith, 1994). The weighted between-item fit statistic (Smith, 1994) is:

$$x^2\ (WB)_i\ =\ (\frac{1}{J-1})\,(\frac{\sum_{j=1}^{J}\left(\sum_{n\epsilon j}^{N_j} x_{ni}-\sum_{n\epsilon j}^{N_j} p_{ni}\right)^2}{\sum_{j=1}^{J}\sum_{n\epsilon j}^{N_j}(p_{ni}\,(1-p_{ni}))})\,. \qquad [11]$$

Both can be transformed through the cube-root transformation (Wright, 1980)

$$t = (MS^{1/3} - 1) (3/S) + (S/3) \quad [12]$$

into a standardized value where S is the SD of $MS\ (UB)_i$ or $MS\ (WB)_i$ (Smith, 1994). The SD of these *MS* is estimated by:

$$S = [\, 2/(J - 1) \,]^{1/2} . \quad [13]$$

Since the simulated measurement disturbances were systematic error, the between-item fit statistics calculated through the above formulas (i.e., Formulas [10]-[13]) should be used in the power analysis (i.e., Research question 3) as suggested by Smith (1994).

Summary

The goal of this simulation study was to examine the Type I error rates and power associated with total item fit statistics in the Rasch dichotomous model. Person ability and item difficulty estimates were first created to obtain response patterns for analyses. It is statistically reasonable to make the assumption that test takers' ability levels follow a normal distribution unless a particular group of individuals is sampled. Furthermore, Rasch models define items that range between -2 and +2 logits as "fitting" (Bond & Fox, 2012). Person abilities were randomly sampled from a normal distribution *N* (0, 1) and item difficulties were randomly sampled from a uniform distribution *U* (-2, +2 logits).

For the simulated data, the numbers of items on a test as well as numbers of test takers were manipulated. In an attempt to mirror the situations for large-scale testing programs, the number of test items were 30, 50, and 70; and the number of examinees were 5,000, 10,000, 25,000, and 50,000. This translated into a combination of 12 crossed conditions to examine the MNSQ values, the ZSTD values, as well as the associated

Type I error rates for the approximate critical values for each condition. Also, comparisons across the conditions that match were examined for general trends and/or patterns.

For a testing situation, various proportions of all items (i.e., 4%, 10%, 20%, and 40%) were manipulated to exhibit moderate to large uniform DIF as defined "C" category by the ETS guideline (Zieky, 2003). The objective was to investigate the power of these critical values to detect these measurement disturbances. Results from the simulated data and simulated measurement disturbances may provide a glimpse on how the total and between-item fit statistics perform in detecting item bias. Researchers and practitioners may find the information useful for item review and instrument validation.

In summary, the simulation study using the original data consists of a fully crossed design with 12 cells: 3 levels of item numbers (i.e., 30, 50, and 70) x 4 sample sizes (5,000, 10,000, 25,000, and 50,000 test-taking students). The investigation on the ability of the chosen indices to detect disturbances consisted of two additional conditions (4 levels of items with a two-group uniform moderate to large DIF, i.e., 4%, 10%, 20% and 40%; 4 magnitudes of moderate to large DIF, i.e., 0.35, 0.45, 0.55 and 0.65 logit units), and this resulted in a fully crossed design with 192 cells: 12 x 16. A total of one thousand replications were run for each design cell as described by Efron (1979) and Wolfe (2013).

CHAPTER 4

RESULTS

This chapter reports the results for the three focal research questions. Research Question 1 examined sensitivity of total item fit statistic (i.e., INFIT mean square and standardized values, OUTFIT mean square and standardized values) in large-scale testing situations under the Rasch dichotomous framework. To determine differences in these four fit indices, results for the mean square (MNSQ) and standardized (ZSTD) values of the two versions of item fit statistics (i.e., INFIT and OUTFIT) are provided.

Research Question 2 then investigated the Type I error rates associated with the rule-of-thumb critical values (i.e., 0.7 and 1.3 for MNSQ values and −2.0 to +2.0 for ZSTD values in general, Wright & Linacre, 1994; for high-stakes situations, 0.8 and 1.2 as suggested, Wright & Linacre, 1994; and for large samples, −3.0 to +3.0, Iramaneerat et al., 2007). Specifically, the fit statistics for these two questions were total item fit indices.

Subsequently, Research Question 3 used the true positive hit rates to evaluate the power of the obtained critical values in detecting simulated item misfit in the previously stated testing conditions, and the 70% threshold was considered acceptable power (González-Romá et al., 2006). As the false positive rates (i.e., Type I error rates, items that are good-performing but flagged as exhibiting DIF) were high in many situations, conditions with well-performing items falsely identified as misfitting were examined and discussed. For this question, between-item fit statistics were used instead of the total

item fit indices due to their sensitivity to systematic measurement disturbances (Smith, 1994).

Item Fit Statistics

Total item fit information was calculated for each simulated data set. INFIT and OUTFIT descriptive summary, including item mean, standard deviation, minimum and maximum values were calculated across 1,000 replications for each of the 12 study conditions. Descriptive information on the mean square values is presented below in Table 4.1 and the descriptive information on the standardized values is listed in Table 4.2.

Across the replications, summary information (see Table 4.1 below) indicated that the weighted and the unweighted mean square average values were extremely consistent across sample size and test length conditions. Almost all the average MNSQ values reached the expected values of 1.0 with a standard deviation value close to 0.0. This was expected and suggested that the simulation was working as intended.

Table 4.1 *Descriptive Information on Mean Square Statistics*

	INFIT				OUTFIT			
	MEAN	SD	MIN.	MAX.	MEAN	SD	MIN.	MAX.
30 Items								
5,000 Test Takers								
Mean	0.999	0.014	0.971	1.028	1.000	0.036	0.931	1.088
SD	0.001	0.002	0.007	0.008	0.004	0.006	0.017	0.030
Min.	0.996	0.008	0.950	1.010	0.986	0.021	0.860	1.030
Max.	1.001	0.020	0.990	1.060	1.016	0.059	0.970	1.260
10,000 Test Takers								
Mean	0.999	0.010	0.979	1.020	1.001	0.025	0.954	1.064
SD	0.001	0.001	0.005	0.006	0.003	0.004	0.012	0.022
Min.	0.997	0.007	0.960	1.010	0.991	0.014	0.910	1.020
Max.	1.001	0.014	0.990	1.040	1.012	0.044	0.980	1.200
25,000 Test Takers								
Mean	0.999	0.007	0.987	1.012	1.000	0.015	0.972	1.039
SD	0.001	0.001	0.005	0.004	0.002	0.002	0.007	0.014

Min.	0.997	0.004	0.970	1.010	0.994	0.009	0.940	1.010
Max.	1.001	0.010	0.990	1.030	1.005	0.024	0.990	1.100
50,000 Test Takers								
Mean	0.999	0.005	0.990	1.010	1.000	0.011	0.980	1.028
SD	0.001	0.001	0.001	0.002	0.001	0.002	0.005	0.009
Min.	0.998	0.002	0.980	1.000	0.996	0.006	0.960	1.010
Max.	1.001	0.008	0.990	1.020	1.004	0.018	0.990	1.070
50 Items								
5,000 Test Takers								
Mean	0.999	0.013	0.971	1.028	0.999	0.029	0.936	1.080
SD	0.001	0.001	0.006	0.006	0.003	0.004	0.014	0.028
Min.	0.997	0.009	0.950	1.020	0.990	0.017	0.860	1.030
Max.	1.001	0.018	0.980	1.060	1.013	0.050	0.970	1.240
10,000 Test Takers								
Mean	0.999	0.009	0.979	1.020	0.999	0.020	0.955	1.052
SD	0.000	0.001	0.004	0.005	0.002	0.002	0.010	0.017
Min.	0.997	0.007	0.960	1.010	0.994	0.012	0.900	1.020
Max.	1.001	0.014	0.990	1.040	1.005	0.029	0.970	1.130
25,000 Test Takers								
Mean	0.999	0.006	0.988	1.012	1.000	0.014	0.970	1.037
SD	0.000	0.001	0.004	0.004	0.001	0.002	0.008	0.011
Min.	0.998	0.004	0.980	1.010	0.996	0.009	0.940	1.010
Max.	1.001	0.009	0.990	1.030	1.004	0.020	0.980	1.090
50,000 Test Takers								
Mean	0.999	0.005	0.990	1.010	0.999	0.010	0.978	1.025
SD	0.000	0.001	0.001	0.001	0.001	0.001	0.005	0.008
Min.	0.997	0.003	0.980	1.000	0.997	0.007	0.960	1.010
Max.	1.000	0.007	0.990	1.020	1.003	0.014	0.990	1.060
70 Items								
5,000 Test Takers								
Mean	0.999	0.013	0.970	1.029	0.999	0.028	0.935	1.081
SD	0.001	0.001	0.006	0.006	0.002	0.003	0.014	0.029
Min.	0.997	0.009	0.950	1.020	0.994	0.020	0.880	1.040
Max.	1.001	0.017	0.980	1.060	1.007	0.044	0.960	1.270
10,000 Test Takers								
Mean	0.999	0.009	0.978	1.020	0.999	0.021	0.951	1.059
SD	0.000	0.001	0.004	0.004	0.001	0.002	0.011	0.019
Min.	0.998	0.007	0.970	1.010	0.994	0.015	0.890	1.030
Max.	1.000	0.012	0.990	1.040	1.005	0.031	0.970	1.160
25,000 Test Takers								
Mean	0.999	0.006	0.987	1.012	0.999	0.013	0.970	1.035
SD	0.000	0.001	0.005	0.004	0.001	0.001	0.007	0.010
Min.	0.998	0.005	0.970	1.010	0.997	0.009	0.940	1.010
Max.	1.000	0.008	0.990	1.020	1.003	0.017	0.980	1.080
50,000 Test Takers								
Mean	0.999	0.005	0.990	1.010	0.999	0.009	0.978	1.025

SD	0.000	0.001	0.001	0.001	0.001	0.001	0.005	0.007
Min.	0.998	0.003	0.980	1.000	0.997	0.006	0.960	1.010
Max.	1.000	0.006	0.990	1.010	1.002	0.013	0.990	1.060

The average OUTFIT MNSQ values were slightly more varied, yielding larger standard deviation values in comparison to their INFIT counterparts (e.g., mean SD of 0.00183 OUTFIT vs 0.00058 INFIT for SD values across all 12 conditions); however, the differences were small enough to be considered negligible. Furthermore, variability across the OUTFIT MNSQ mean values slightly decreased with as the number of test takers and/or items increased. For example, with a test of 30 items, the standard deviation value of the mean values was 0.004 for 5,000 test takers, 0.003 for 10,000 test takers, to 0.002 for 25,000 and 0.001 for 50,000 test takers. Similar patterns were observed for a test of 50 and 70 items holding constant the number of test items. Overall, the number of test takers and test items appeared to exert a small effect on the mean of the weighted and unweighted mean square values.

The standard deviation of the MNSQ values appeared to differ for weighted and unweighted statistics. Consistent with Smith et al. (1998), here, mean SD values of the OUTFIT statistic were almost doubled or tripled of those of the INFIT statistic (e.g., for 30 items with 5,000 test takers, the mean INFIT SD was 0.014 and OUTFIT SD was 0.036), suggesting that Type I error rate and critical values cannot be the same for the MNSQ values of these two indices. Also, it appeared that the mean SD decreased as the number of test takers increased. For example, for the INFIT statistic with a 50-item test, the mean SD values declined from 0.013 with 5,000 test takers to 0.005 with 50,000 test takers. Similarly, for the OUTFIT statistic for the same test, the SD values dropped from 0.029 with 5,000 testing students to 0.010 with 50,000 students. The length of the test;

however, did not appear to have much impact on the SD of these two item fit indices. Values were relatively constant, with INFIT values from 0.007, 0.006 to 0.006 for 25,000 test takers and from 0.015, 0.014, to 0.013 for 25,000 test takers in OUTFIT across the three test length conditions.

Similar patterns were observed with the range of mean INFIT and OUTFIT MNSQ indices. Even though the range of these two indices was slightly different, the OUTFIT statistic reported slightly wider ranges (i.e., 0.157 for OUTFIT and 0.057 for INFIT for a testing situation of 30 items and 5,000 test takers). Overall, the mean range of both indices decreased as the number of test items increased. For example, in a 70-item test, the OUTFIT statistic (average) range declined from 0.146 (1.081 – 0.935), 0.108 (1.059 – 0.951), 0.065 (1.035 – 0.970) to 0.047 (1.025 – 0.978) as the number of test takers increased from 5,000 to 50,000. This finding was also consistent with the previous Smith et al. (1998) study, in which the mean range for the INFIT and OUTFIT statistic decreased as the number of test takers increased. The mean range values of the OUTFIT were less than half of those INFIT values (i.e., the INFIT average range differences were 0.059, 0.042, 0.025 and 0.020 for a test with 70 items).

In comparison to the MNSQ values, ZSTD values were a bit different (see Table 4.2). The average *t*-transformation standardized values were close to their expected values of 0.0 with a standard deviation close to 1.0. The OUTFIT ZSTD values were closer to the expected values than their INFIT counterparts, but exhibited more variation, with larger SD values. For instance, in a 30-item test, the SD of mean for the weighted statistic was approximately 0.033 regardless of the number of test takers, while this the unweighted SD of the mean was approximately 0.097, almost three-fold. Overall, the

numbers of test items and test takers did not appear to have an influence on the ZSTD values.

Table 4.2 *Descriptive Information on Standardized Statistics*

	INFIT				OUTFIT			
	MEAN	SD	MIN.	MAX.	MEAN	SD	MIN.	MAX.
30 Items								
5,000 Test Takers								
Mean	-0.077	0.796	-1.826	1.536	-0.021	0.927	-1.858	2.009
SD	0.033	0.109	0.466	0.398	0.099	0.131	0.406	0.530
Min.	-0.177	0.447	-3.650	0.670	-0.350	0.587	-3.480	0.650
Max.	0.018	1.279	-0.820	3.160	0.315	1.509	-0.960	4.670
10,000 Test Takers								
Mean	-0.102	0.838	-1.969	1.562	-0.019	0.949	-1.929	2.018
SD	0.034	0.113	0.473	0.400	0.100	0.128	0.454	0.514
Min.	-0.239	0.521	-4.890	0.610	-0.327	0.571	-3.790	0.770
Max.	0.002	1.210	-1.000	4.410	0.345	1.317	-0.810	4.580
25,000 Test Takers								
Mean	-0.177	0.880	-2.118	1.532	-0.066	0.970	-2.057	1.952
SD	0.033	0.120	0.477	0.393	0.097	0.124	0.451	0.487
Min.	-0.293	0.490	-4.900	0.650	-0.377	0.649	-3.600	0.780
Max.	-0.087	1.315	-1.060	2.970	0.224	1.411	-0.990	3.650
50,000 Test Takers								
Mean	-0.196	0.936	-2.280	1.594	-0.077	1.014	-2.170	2.001
SD	0.032	0.127	0.496	0.407	0.092	0.132	0.474	0.497
Min.	-0.289	0.592	-4.420	0.690	-0.414	0.565	-3.920	0.760
Max.	-0.088	1.487	-1.260	3.280	0.253	1.524	-1.060	3.840
50 Items								
5,000 Test Takers								
Mean	-0.064	0.809	-2.013	1.792	-0.034	0.913	-2.026	2.201
SD	0.024	0.085	0.428	0.398	0.068	0.097	0.401	0.553
Min.	-0.145	0.555	-4.160	0.940	-0.254	0.590	-3.800	1.030
Max.	0.022	1.095	-1.090	3.800	0.301	1.311	-0.950	4.750
10,000 Test Takers								
Mean	-0.093	0.835	-2.098	1.822	-0.055	0.927	-2.113	2.169
SD	0.024	0.089	0.449	0.424	0.062	0.098	0.423	0.530
Min.	-0.180	0.589	-4.040	0.950	-0.235	0.614	-3.810	1.070
Max.	-0.017	1.186	-1.050	4.030	0.140	1.250	-1.220	5.300
25,000 Test Takers								
Mean	-0.120	0.836	-2.164	1.755	-0.063	0.941	-2.168	2.133
SD	0.026	0.090	0.454	0.407	0.069	0.094	0.426	0.462
Min.	-0.201	0.575	-3.790	0.650	-0.276	0.628	-3.880	1.030
Max.	-0.026	1.164	-1.150	3.350	0.166	1.256	-1.190	3.790
50,000 Test Takers								

Mean	-0.248	0.826	-2.316	1.531	-0.128	0.940	-2.281	2.007
SD	0.026	0.090	0.493	0.377	0.070	0.098	0.446	0.465
Min.	-0.341	0.559	-4.730	0.720	-0.368	0.627	-4.310	0.790
Max.	-0.168	1.068	-1.270	3.120	0.106	1.246	-1.240	3.820
70 Items								
5,000 Test Takers								
Mean	-0.058	0.803	-2.108	1.923	-0.034	0.908	-2.133	2.343
SD	0.025	0.073	0.428	0.423	0.056	0.086	0.388	0.584
Min.	-0.135	0.552	-4.140	0.950	-0.201	0.671	-3.900	1.200
Max.	0.013	1.030	-1.060	4.300	0.163	1.248	-1.230	5.130
10,000 Test Takers								
Mean	-0.098	0.776	-2.096	1.797	-0.058	0.899	-2.134	2.236
SD	0.021	0.070	0.412	0.401	0.056	0.080	0.378	0.510
Min.	-0.160	0.587	-3.940	0.860	-0.227	0.670	-3.720	1.190
Max.	-0.032	1.022	-1.250	3.850	0.120	1.142	-1.230	4.240
25,000 Test Takers								
Mean	-0.137	0.795	-2.200	1.781	-0.082	0.909	-2.233	2.180
SD	0.020	0.075	0.433	0.398	0.057	0.081	0.387	0.471
Min.	-0.205	0.562	-3.930	0.990	-0.262	0.702	-3.660	1.220
Max.	-0.076	1.026	-1.320	3.760	0.113	1.170	-1.190	4.560
50,000 Test Takers								
Mean	-0.206	0.842	-2.376	1.796	-0.144	0.933	-2.387	2.122
SD	0.024	0.077	0.437	0.409	0.053	0.080	0.424	0.441
Min.	-0.277	0.628	-4.530	0.940	-0.305	0.686	-4.340	1.180
Max.	-0.090	1.127	-1.290	3.580	0.015	1.226	-1.290	3.950

Again, the standard deviation of the unweighted ZSTD values were closer to the expected values than for the weighted ZSTD values; however, both indices were not greatly influenced by the sample size or test length. For instance, the SD value of the average INFIT statistic was 0.809, 0.835, 0.836, and 0.826 for a 50-item test across the different sample sizes included here, and this value for the OUTFIT statistic was 0.913, 0.927, 0.941, and 0.994 for the same testing situations. For a group of 5,000 examinees, the standard deviation of the mean INFIT statistic were 0.796 (30 items), 0.809 (50 items) and 0.803 (70 items) whereas the OUTFIT values were 0.927, 0.913, and 0.908 for the same testing scenarios. Furthermore, for 30 items and 5,000 test takers, the mean and SD of the average OUTFIT ZSTD were −0.021 and 0.927. Even with 70 items and

50,000 test takers, these values were only −0.144 and 0.933. These showed that the Bond and Fox (2012) suggestion that all items would show misfit if sample sizes were large enough does not hold.

The patterns were also similar in the ranges of mean weighted and unweighted ZSTD values. Overall, the OUTFIT statistic had a wider range than the INFIT statistic, but both indices did not appear to be influenced by sample size or test length.

Type I Error Rates and Critical Values

To investigate the Type I error rates associated with the INFIT and OUTFIT statistics, the percentage of values exceeding the recommended cut-off values were calculated. Specifically, the percentage of MNSQ values exceeding ranges of 0.7 and 1.3, 0.8 and 1.2, as well as 0.9 and 1.1 while the percentage of ZSTD values exceeding ±2.0, ±3.0, and ±4.0 were calculated following suggestions from previous research (Smith et al.,1998).

Results from the weighted MNSQ showed values greater than the commonly used rule-of-thumb for detecting measurement disturbances did not occur in any of the 12 large-scale testing situations. For the unweighted MNSQ values, when the sample size of test takers was 5,000, values greater than 1.1 occurred at a rate of approximately 1.00% in a 30-item test. This number declined to around a third of the rate (0.34%) for a 50-item test, and then further declined to 0.24% for a test of 70 items. If the cut-off value of 1.2 was used, values greater than 1.2 occurred around 0.01% for 5,000 test takers. Both cut-off values are liberal given the customary error rate of 2.5% for a one-tail test with a symmetrical distribution.

Consistent with results from Smith et al. (1998), the number of INFIT MNSQ values greater than the traditionally used critical values were too small to determine the effect of test length. Furthermore, the OUTFIT MNSQ statistic were also too small to determine the influence of test length, even though in some testing situations there might be a small effect.

As results from the MNSQ values provided very low occurrences of Type I error rates, percentage of ZSTD statistic greater than the common cut-off values were also calculated. Results for ZSTD values are presented in Table 4.3.

Table 4.3 *Percentage of Cases above Given Critical Values across 1,000 Replications*

	INFIT ZSTD				OUTFIT ZSTD			
	5,000	10,000	25,000	50,000	5,000	10,000	25,000	50,000
30 Items								
> 4.0	0.000	0.003	0.000	0.000	0.017	0.007	0.000	0.000
> 3.0	0.010	0.010	0.000	0.013	0.143	0.137	0.120	0.123
> 2.0	0.417	0.470	0.410	0.550	1.920	1.910	1.643	1.910
< −2.0	1.183	1.723	2.467	3.460	1.287	1.510	2.250	2.770
< −3.0	0.056	0.077	0.163	0.290	0.033	0.100	0.117	0.177
< −4.0	0.000	0.003	0.010	0.010	0.000	0.000	0.000	0.000
50 Items								
> 4.0	0.000	0.004	0.000	0.000	0.018	0.010	0.000	0.000
> 3.0	0.012	0.024	0.012	0.006	0.184	0.136	0.094	0.062
> 2.0	0.586	0.702	0.568	0.210	1.722	1.568	1.568	1.246
< −2.0	1.136	1.448	1.712	2.442	1.192	1.536	1.780	2.294
< −3.0	0.048	0.074	0.086	0.194	0.042	0.070	0.098	0.138
< −4.0	0.002	0.002	0.000	0.012	0.000	0.000	0.000	0.008
70 Items								
> 4.0	0.001	0.000	0.000	0.000	0.027	0.009	0.001	0.000
> 3.0	0.023	0.016	0.013	0.011	0.201	0.116	0.089	0.063
> 2.0	0.621	0.426	0.389	0.437	1.689	1.410	1.283	1.140
< −2.0	1.146	1.057	1.400	2.107	1.256	1.224	1.650	2.234
< −3.0	0.051	0.030	0.077	0.123	0.037	0.031	0.073	0.133
< −4.0	0.001	0.000	0.000	0.003	0.000	0.000	0.000	0.004

Unlike the MNSQ values, result on the *t*-transformation statistic suggested that the traditional ±2.0 appeared to be effective in serving as cut-off values, even in large-

scale testing conditions. In a test of 30 items, the error rates were close to or below the .05 level. For instance, if 5,000 students took this test, 1.60% (i.e., 0.417% + 1.183%) of items may be mis-classified as non-fitting to the model if the INFIT ZSTD values were used whereas 3.42% of items may be mis-identified if OUTFIT ZSTD statistic were applied. Furthermore, all the 12 testing conditions yielded slightly higher error rates if OUTFIT ZSTD were used instead of the INFIT ZSTD values; however, all values were still within the traditionally accepted range. Under longer testing situations (e.g., 50 and 70 items), again the OUTFIT standardized statistic had a slightly higher error rates than the INFIT standardized values, but all were still within the acceptable error level. Even though these values are slightly lower than the traditional .05 error rate, using the values of ±1.0 may result in error rates that exceed the traditional cutoff, especially when the number of test takers is large.

True Positive Hit Rates, Power Analysis and False Positive Rates

The results from the above simulated data sets showed that almost all the items were correctly identified as good-fitting items. However, the item fit indices used to investigate INFIT and OUTFIT were total item fit statistics. For systematic errors as in DIF items, the total fit statistic may report low power and the between-fit item statistic should be used instead (Smith, 1994). Table 30.4 in Winsteps® does not report results on mean square and standardized values of the INFIT statistic; and therefore, only the results of the OUTFIT indices were summarized.

Hit rates and power analysis. Results from across the replications indicated that both MNSQ and ZSTD indices could correctly identify items with moderate to large

magnitude of DIF (i.e., ETS level "C", Zieky, 2003). Here, MNSQ values slightly outperformed ZSTD values in true positive rates (i.e., power).

Regardless of the number of test takers, test length, the percentage of DIF items, and the magnitude of DIF, majority of the unweighted between-fit statistics (97%, *n*=745 from 768 estimates including 3 test lengths x 4 numbers of test takers x 4 different percentages of DIF items x 4 conditions of DIF magnitude x 4 cut-off values) successfully flagged the items with at least 99% accuracy. Only three out of the remaining 23 individual estimates obtained an accuracy rate of less than 90%; and all lower hit rates were obtained in the 0.35 logits DIF conditions. When the item difference was 0.35 logit more difficult for one group of 2,500 students than the other group, the hit rates dropped to lower than 85% (using ±3.0 as the cut-off value) for a test with 40% DIF items. Specifically, these hit rates were 79%, 83%, and 84% for a test of 30, 50, and 70 items respectively. If ±2.0 was used instead as the cutoff, the hit rates were 93%, 96%, and 96%, only slightly lower than the 99% or 100% accuracy rates if their MNSQ counterpart values were used. Furthermore, the lowest hit rate was 79% (i.e., a 30-item test with 5,000 test takers and 40% of 0.35 logits DIF item using ±3.0 as the cut-off value), which was still higher than the threshold of 70%. It was also evident that as the magnitude of DIF increased, the power increased. In other words, the percentage of correctly identifying misfitting items increased when the magnitude of DIF changed from 0.35 to 0.65 logits (see Figure 4.1 for using the ±3.0 as the ZSTD cut-off). However, the increase was not reliant on a larger amount of DIF as the true positive rates were high overall across all conditions tested, especially for the other testing situations and other value estimates. The power curve was basically a flat straight line in other situations.

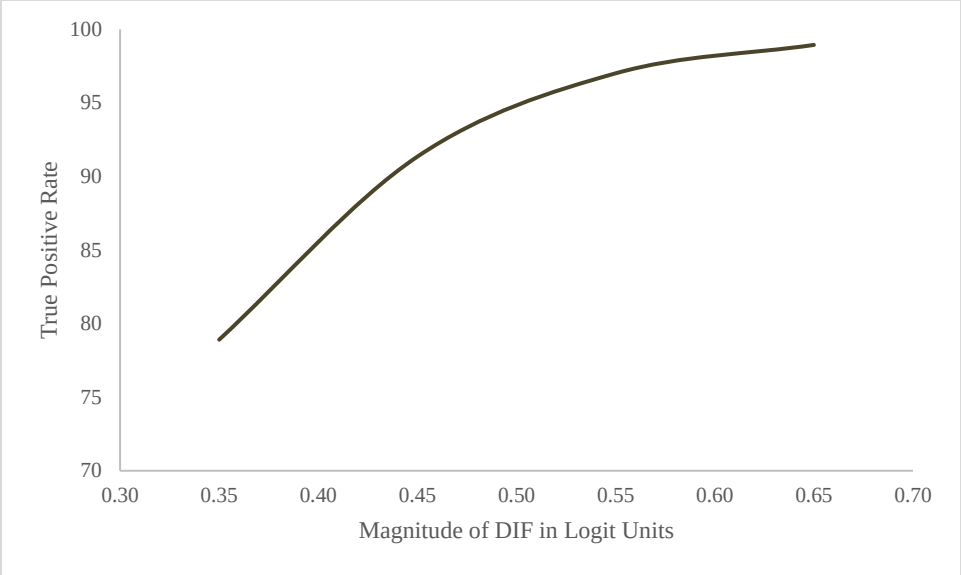

Figure 4.1. Power Curve for a 30-item Test with 5,000 Test Takers and 12 (40%) DIF Items

False positive rates. As correctly identifying misfitting items is crucial, correctly identifying items that fit the model is also important. A false positive rate (i.e., Type I error rate) indicates the percentage of model-fitting items being mis-classified as non-fitting. While power was high for the study; however, results suggested that the number of false positive cases may also be high. Tables 4.4-4.7 note the percentage of items that were incorrectly identified as ill-functioning, with magnitudes of DIF items ranging from 0.35, 0.45, 0.55 to 0.65 logit units.

In general, for the OUTFIT statistics reported by Winsteps®, the ZSTD values greatly outperformed the MNSQ values in mis-classifying items without DIF when the percentage of DIF items constituted less than 20% of the test, and the number of test takers were fewer than 50,000 students. In a situation where 20% or more of DIF items were present and 50,000 students took the test, the ZSTD statistic still slightly outperformed their MNSQ counterparts, even though the error rates were still high.

Overall, when considering the DIF magnitude situations, the false positive rates increased as the percentage of DIF items increased and the number of test takers increased, for both MNSQ and ZSTD values. Nevertheless, MNSQ values experienced a smaller increase in comparison to their ZSTD counterparts since the MNSQ indices were already high.

Also, results suggested that the number of test takers exerted a slightly greater impact than the proportion of DIF items. For instance, when DIF items differed by 0.35 logits in a 30-item test, the false positive rates of ZSTD (±2.0) went from 2.87%, 5.00%, 15.31% to 69.36%, with a latter value more than doubled the previous condition as the percentage of DIF items increased. For the same index, the rates increased to 9.02% (with 4% DIF items), 34.94% (with 10% DIF items), 88.38% (with 20% DIF items) and 99.99% (with 40% DIF items) when the number of test-taking students rose from 5,000 to 50,000. Similar patterns were observed in 50-item and 70-item tests as well as different magnitude of DIF items. This was somewhat different from the previous results (from Tables 4.1-4.2) in which the number of test takers had a small influence on ZSTD indices when the items were all well-fitting to the model.

Across the different DIF magnitude situations, the changes of MNSQ values appeared to be smaller than the changes of ZSTD values, especially when the percentage of DIF items was low and the number of test takers were at the lower end of the conditions. For example, a 50-item test with 5 (10%) DIF items, yielded false positive rates were 5.22% (± 2.0) and 0.46% (± 3.0), 7.09% (± 2.0) and 0.69% (± 3.0), 9.54% (± 2.0) and 1.07% (± 3.0), as well as 12.60% (± 2.0) and 1.64% (± 3.0) for each of the different magnitude testing conditions increased from 0.35 to 0.65 logits. In the

meantime, the MNSQ values remained stable for these conditions, ranging between 85.25% to 86.91% when (0.7, 1.3) were used as the cut-off values and between 90.41% to 91.44% when the cut-off values of (0.8, 1.2) were applied. Again, the changes of ZSTD values were more apparent along with the increase in DIF item presence and the number of test takers. Results suggested that the increase of error greatly impacted the performance of fit indices. It appeared that the additional amount of error introduced by more test takers and more items were distributed to non-DIF items. This was more evident in situations of 40% DIF items. When a large proportion of DIF items exhibited item DIF, error appeared to be more evenly distributed to all items in all test length and test taker situations. This resulted in very high false positive rates, even for ZSTD indices using ±2.0 or ±3.0 as the cut-off values.

The following graphs (Figures 4.2-4.4) showcase the false positive rates (using ±2.0 for ZSTD values) for different magnitude of DIF as the percentage of DIF items included on a test changed. When the proportion was small and the number of test takers was no more than 10,000, the error rates were low across moderate and large DIF conditions. When the proportion of DIF items spiked to 40%, however, the error rates were high (at least 60%) regardless of the sample size, test length, and magnitude of DIF. For a test including 20% of items with DIF, when the number of test takers were at least 10,000, the false positive rates accelerated to more than 80% and leveled off to about 100%.

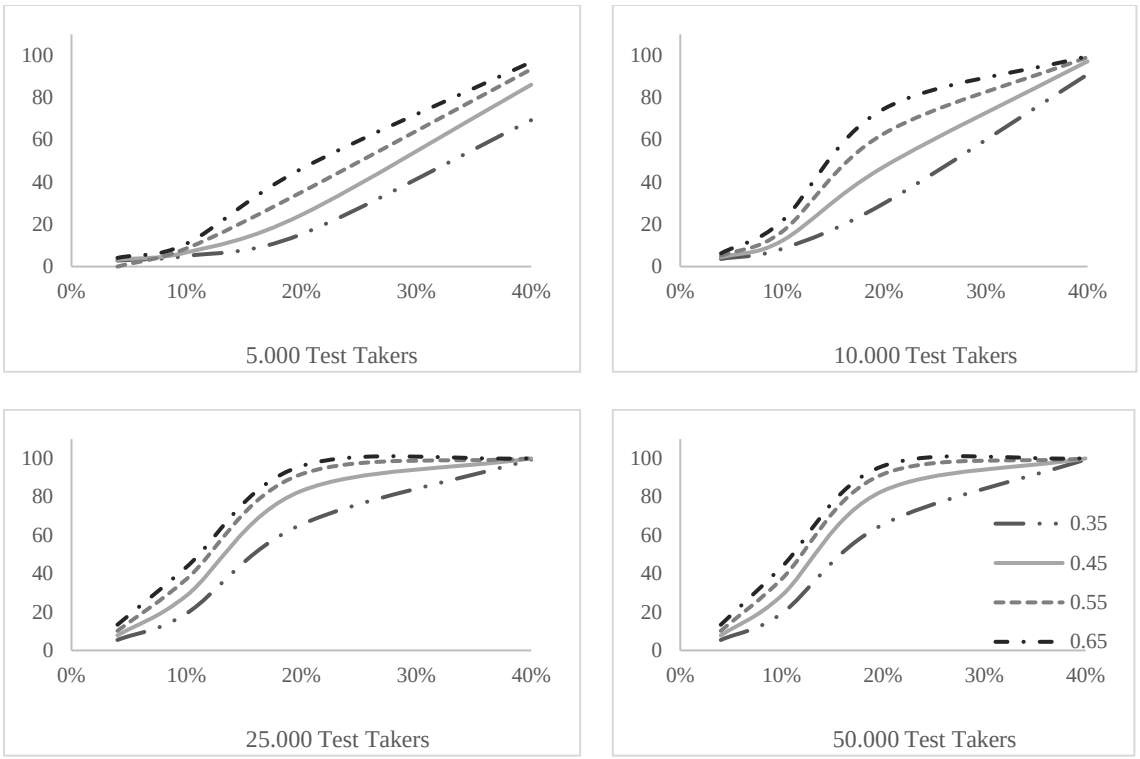

Figure 4.2. False Positive Rates by DIF Magnitude for a 30-item Test

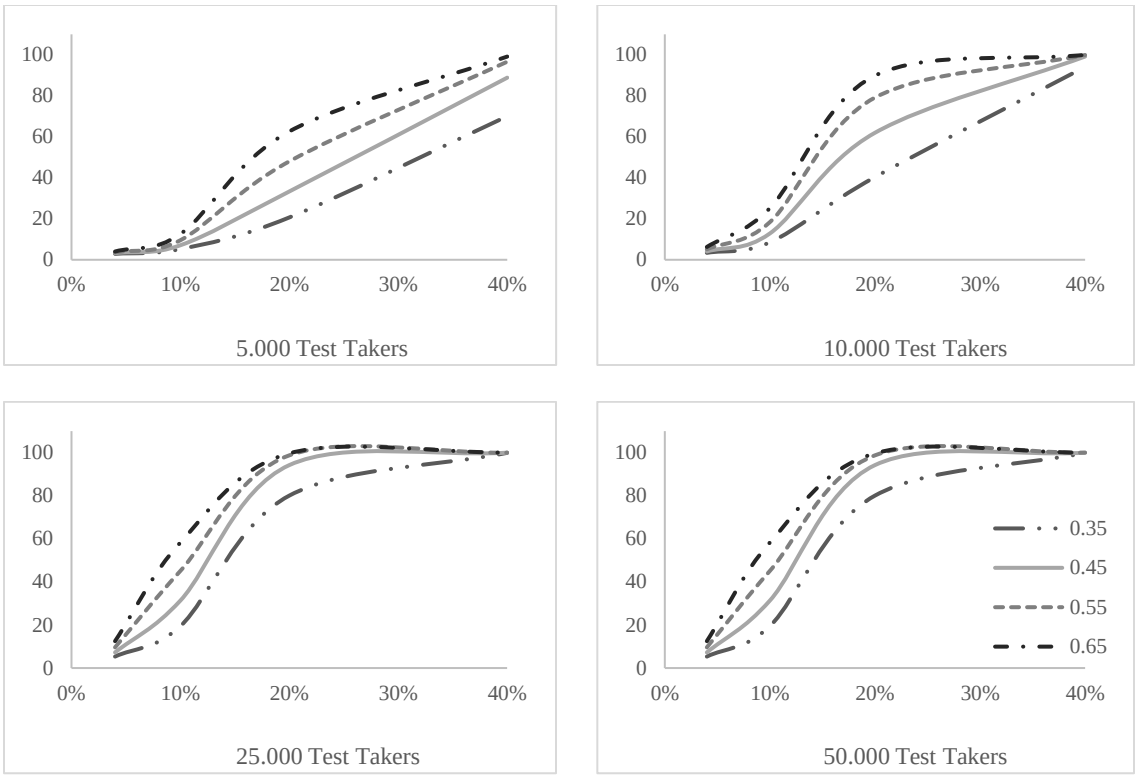

Figure 4.3. False Positive Rates by DIF Magnitude for a 50-item Test

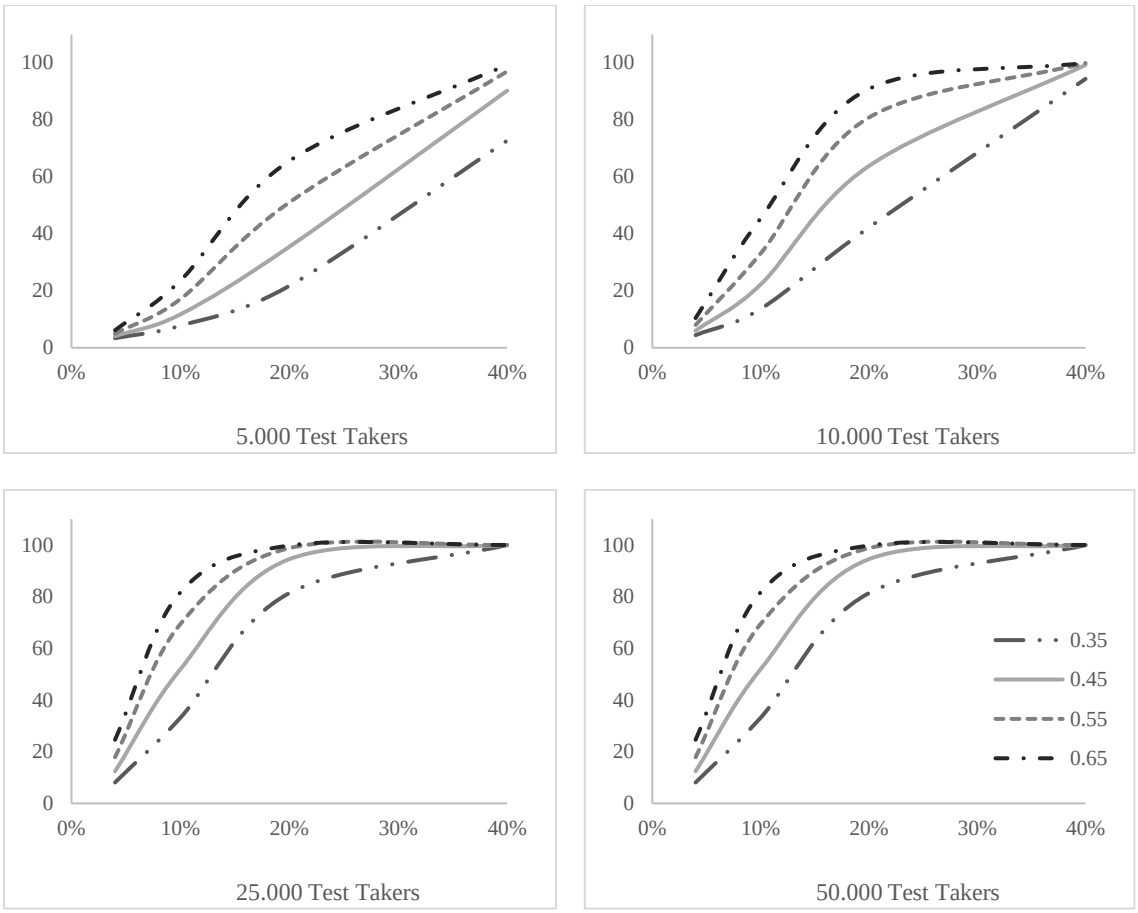

Figure 4.4. False Positive Rates by DIF Magnitude for a 70-item Test

DIF of 0.35 logits. False positive rate results for a moderate DIF difference of 0.35 logits were presented below (see Table 4.4).

Table 4.4 *Percentage of False Positive Cases across 1,000 Replications with a Difference of 0.35 Logits*

Number of DIF Items	Fit Index	Cut-off Values	Number of Test Takers 5,000	10,000	25,000	50,000
30 Items						
1	MSNQ	0.7, 1.3	85.14	84.94	85.28	85.62
		0.8, 1.2	90.28	90.21	90.05	90.56
	ZSTD	±2.0	2.87	3.61	5.44	9.02
		±3.0	0.18	0.29	0.50	1.14
3	MSNQ	0.7, 1.3	84.99	85.94	87.85	91.16
		0.8, 1.2	90.28	90.73	91.89	94.46
	ZSTD	±2.0	5.00	8.41	19.24	34.94

		±3.0	0.35	1.07	4.10	13.22
6	MSNQ	0.7, 1.3	87.17	90.30	96.65	99.14
		0.8, 1.2	91.42	93.50	97.74	99.43
	ZSTD	±2.0	15.31	29.91	65.53	88.38
		±3.0	2.29	7.31	32.27	68.50
12	MSNQ	0.7, 1.3	97.03	99.37	99.96	100.00
		0.8, 1.2	98.09	99.62	99.97	100.00
	ZSTD	±2.0	69.36	90.99	99.39	99.99
		±3.0	36.13	72.66	96.65	99.84
50 Items						
2	MSNQ	0.7, 1.3	85.24	85.25	85.47	85.91
		0.8, 1.2	90.24	90.23	90.41	90.61
	ZSTD	±2.0	2.89	3.38	5.33	8.14
		±3.0	0.20	0.19	0.43	0.87
5	MSNQ	0.7, 1.3	85.25	85.72	87.90	91.94
		0.8, 1.2	90.41	90.55	91.98	94.66
	ZSTD	±2.0	5.22	8.52	19.46	38.28
		±3.0	0.46	0.86	3.39	10.45
10	MSNQ	0.7, 1.3	88.89	92.38	98.43	99.91
		0.8, 1.2	92.69	94.90	98.98	99.92
	ZSTD	±2.0	20.66	40.35	80.06	97.21
		±3.0	3.85	11.56	47.74	85.58
20	MSNQ	0.7, 1.3	97.36	99.74	100.00	100.00
		0.8, 1.2	98.26	99.83	100.00	100.00
	ZSTD	±2.0	70.04	94.10	99.91	100.00
		±3.0	34.44	75.11	99.15	100.00
70 Items						
3	MSNQ	0.7, 1.3	85.35	85.22	85.84	86.85
		0.8, 1.2	90.26	90.45	90.30	91.31
	ZSTD	±2.0	3.37	4.42	8.08	14.86
		±3.0	0.22	0.33	0.81	2.33
7	MSNQ	0.7, 1.3	85.58	86.87	91.09	95.63
		0.8, 1.2	90.52	91.27	94.16	97.30
	ZSTD	±2.0	7.74	13.64	33.12	59.20
		±3.0	0.79	2.00	8.55	26.26
14	MSNQ	0.7, 1.3	88.55	93.06	98.50	99.87
		0.8, 1.2	92.34	95.43	99.07	99.91
	ZSTD	±2.0	21.72	42.20	81.36	97.13
		±3.0	3.99	12.63	50.30	86.17
28	MSNQ	0.7, 1.3	97.52	99.74	100.00	100.00
		0.8, 1.2	98.35	99.82	100.00	100.00
	ZSTD	±2.0	72.69	94.36	99.94	100.00
		±3.0	38.42	77.15	99.03	100.00

Note: Shaded cells show the minimum/maximum values.

A closer look at the results of the moderate DIF (i.e., 0.35 logits) showed that the percentage of false positive rates in MNSQ values ranged from 84.94% to 100.00% with the traditional cutoffs 0.7 and 1.3, and from 90.05% to 100.00% using cut-off values of 0.8 and 1.2. In contrast, ZSTD (±2.0) performed well with a false positive rate approximating the Type I error rate of .05 when a small number of DIF items were present (e.g., 4% of the test) and sample sizes of examinees were smaller than 25,000. For tests with more DIF items and test takers, ZSTD (±3.0) functioned well. Specifically, when the proportion of DIF items reached 10%, using ZSTD (±3.0), the majority of the testing conditions still achieved a .05 Type I error rate.

Error rates exceeded .05 when the number of test takers rose to 50,000. One exception was the Type I error rate with a 70-item test for 25,000 test takers. It was 8.55% (or .0855), slightly above the commonly used .05. The value of ±3.0 still functioned well when the proportion of DIF items reached 20% with 5,000 test takers. As the number of DIF items increased to 40%, or the number of test takers increased, even this cut-off value resulted in a large number of items being wrongly identified as ill-fitting.

As seen from the graphical display below (Figures 4.5-4.7), when ±2.0 was used to identify item misfit for the OUTFIT ZSTD statistic, the false positive rates increased rapidly as the number of items with moderate DIF increased. The increase was notable when the number of DIF items reached higher levels, including 20% and 40% of the test. Basically, the item fit indices flagged majority of the model-fitting items as misfitting (i.e., at least about 70%) regardless of the presence or absence of DIF. For a group of test

takers of 5,000 and 10,000, the change in false positive rates appeared more as a straight line for all three different test lengths. This was somewhat expected.

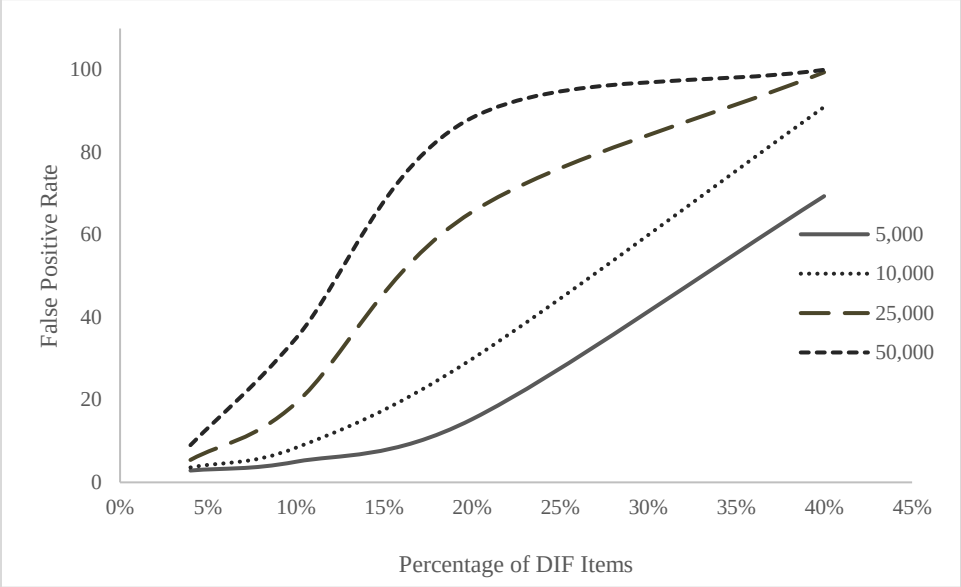

Figure 4.5. False Positive Rates for a 30-item Test with a Difference of 0.35 Logits

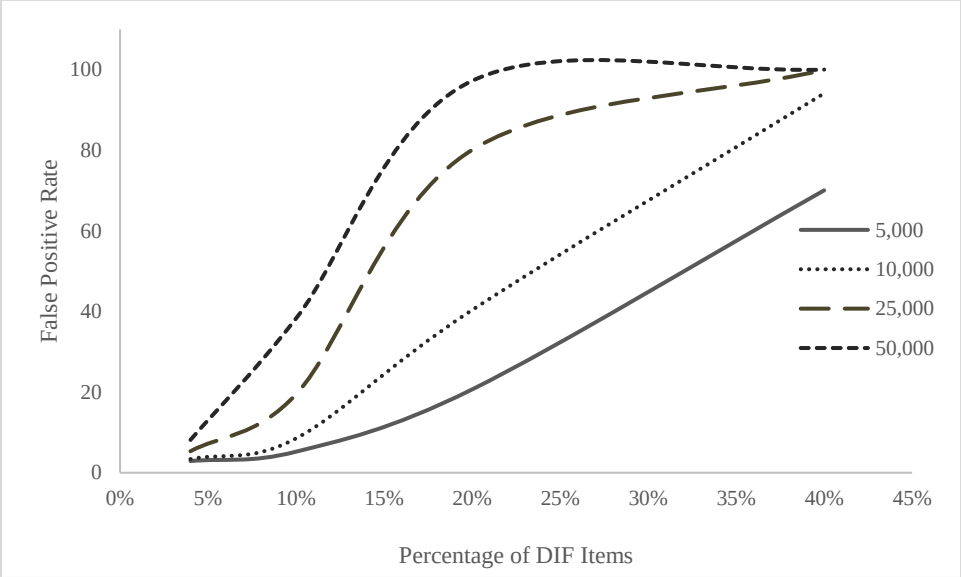

Figure 4.6. False Positive Rates for a 50-item Test with a Difference of 0.35 Logits

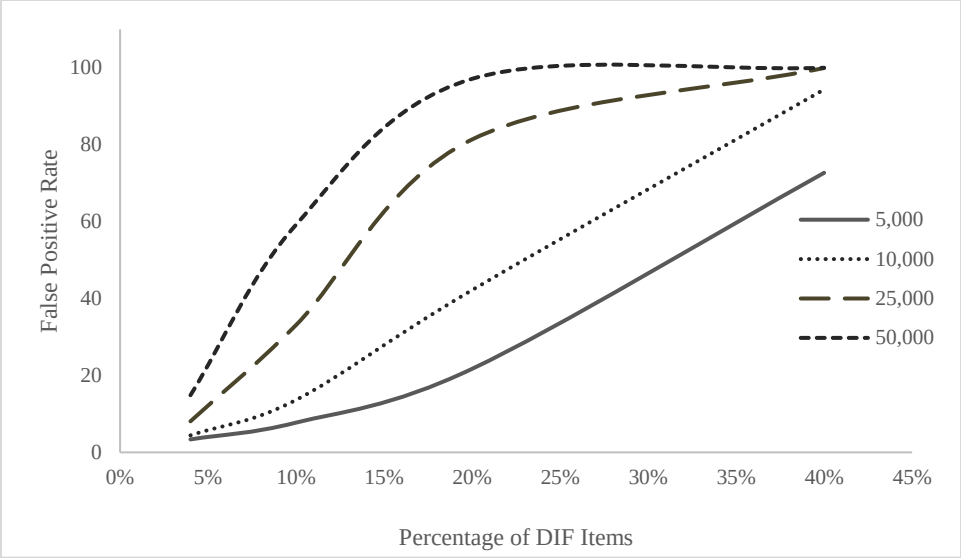

Figure 4.7. False Positive Rates for a 70-item Test with a Difference of 0.35 Logits

DIF of 0.45 logits. False positive rate results for a DIF difference of 0.45 logits were presented as the following.

Table 4.5 *Percentage of False Positive Cases across 1,000 Replications with a Difference of 0.45 Logits*

Number of DIF Items	Fit Index	Cut-off Values	Number of Test Takers 5,000	10,000	25,000	50,000
30 Items						
1	MSNQ	0.7, 1.3	85.26	85.10	85.95	86.77
		0.8, 1.2	90.19	90.27	90.89	91.16
	ZSTD	±2.0	3.23	4.37	7.81	13.43
		±3.0	0.21	0.38	0.75	2.20
3	MSNQ	0.7, 1.3	85.51	86.80	89.72	92.96
		0.8, 1.2	90.54	91.34	93.06	95.25
	ZSTD	±2.0	6.69	12.21	28.51	46.84
		±3.0	0.60	1.93	9.01	24.54
6	MSNQ	0.7, 1.3	89.39	93.87	98.58	99.85
		0.8, 1.2	92.95	96.05	99.05	99.88
	ZSTD	±2.0	24.45	47.30	83.02	96.32
		±3.0	4.87	16.30	56.85	86.04
12	MSNQ	0.7, 1.3	98.93	99.88	100.00	100.00
		0.8, 1.2	99.32	99.92	100.00	100.00
	ZSTD	±2.0	86.14	97.19	99.93	100.00
		±3.0	62.03	89.09	99.35	100.00

50 Items						
2	MSNQ	0.7, 1.3	85.15	85.11	85.58	86.67
		0.8, 1.2	90.22	90.18	90.34	91.50
	ZSTD	±2.0	3.24	4.07	7.24	12.28
		±3.0	0.24	0.27	0.63	1.65
5	MSNQ	0.7, 1.3	85.49	86.94	90.91	95.75
		0.8, 1.2	90.44	91.47	93.85	97.13
	ZSTD	±2.0	7.09	12.85	31.32	58.24
		±3.0	0.69	1.68	7.52	23.51
10	MSNQ	0.7, 1.3	91.08	96.16	99.73	100.00
		0.8, 1.2	94.16	97.49	99.80	100.00
	ZSTD	±2.0	33.34	61.96	94.23	99.75
		±3.0	8.36	26.76	76.46	97.51
20	MSNQ	0.7, 1.3	99.27	99.98	100.00	100.00
		0.8, 1.2	99.53	99.99	100.00	100.00
	ZSTD	±2.0	88.86	99.17	100.00	100.00
		±3.0	63.08	94.16	99.95	100.00
70 Items						
3	MSNQ	0.7, 1.3	85.39	85.57	86.31	88.97
		0.8, 1.2	90.39	90.49	91.13	92.68
	ZSTD	±2.0	4.07	5.99	12.40	23.66
		±3.0	0.30	0.54	1.64	4.99
7	MSNQ	0.7, 1.3	86.45	88.58	94.37	98.02
		0.8, 1.2	91.00	92.55	96.35	98.78
	ZSTD	±2.0	11.72	22.17	51.90	79.03
		±3.0	1.47	4.34	19.87	50.99
14	MSNQ	0.7, 1.3	91.67	96.65	99.74	99.99
		0.8, 1.2	94.45	97.72	99.83	100.00
	ZSTD	±2.0	35.44	63.79	94.51	99.71
		±3.0	9.33	28.80	78.27	97.56
28	MSNQ	0.7, 1.3	99.36	99.97	100.00	100.00
		0.8, 1.2	99.59	99.98	100.00	100.00
	ZSTD	±2.0	90.22	99.24	100.00	100.00
		±3.0	67.08	94.59	99.99	100.00

Note: Shaded cells show the minimum/maximum values.

When the moderate item DIF increased to 0.45 logits, results indicated that the percentage of false positive rates in MNSQ values ranged from 85.10% to 100.00% for (0.7, 1.3) and from 90.18% to 100.00% for (0.8, 1.2). In contrast, ZSTD (±2.0) again performed well with a false positive rate close to the Type I error rate of .05 when a small

number of DIF items were present (e.g., 4% of the whole test) and sample sizes of examinees were smaller than 25,000. For tests with more DIF items and test takers, ZSTD (±3.0) functioned well. Specifically, when the proportion of DIF items reached 10%, majority of the testing conditions still achieved a .05 Type I error rate unless the number of test takers reached 25,000.

The following graphs (Figures 4.8-4.10) suggested that again the false positive rates increased substantially as the number of DIF items increased when ±2.0 was used to identify item misfit for the OUTFIT ZSTD statistic. This jump was apparent when the number of DIF items reached 20% and the false positive rate started to level off for conditions in which DIF items constituted 40% of the test. Basically, the item fit indices indicated at least 85% the well-performing test items as misfitting. For a group of 5,000 test takers, the false positive rates appeared more as a straight line for all three test length conditions.

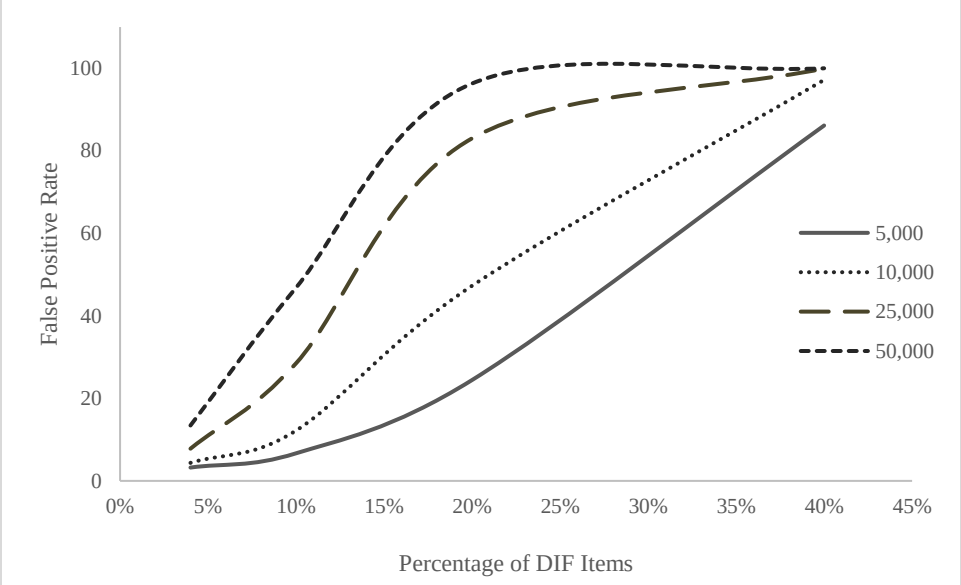

Figure 4.8. False Positive Rates for a 30-item Test with a Difference of 0.45 Logits

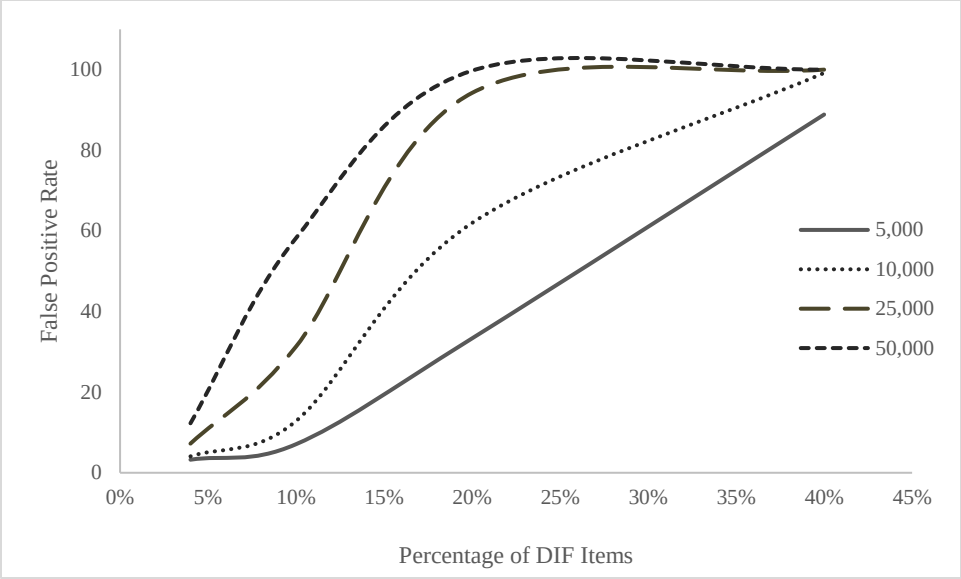

Figure 4.9. False Positive Rates for a 50-item Test with a Difference of 0.45 Logits

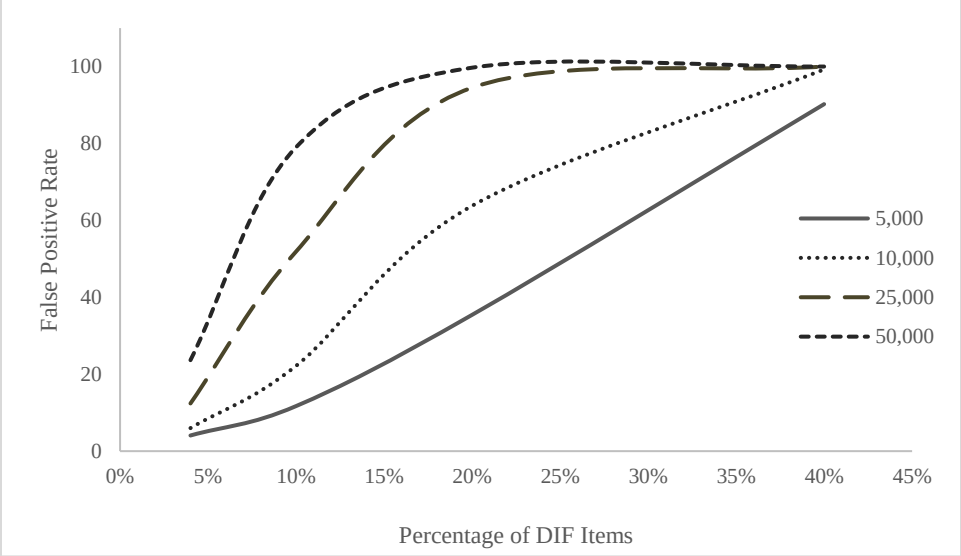

Figure 4.10. False Positive Rates for a 70-item Test with a Difference of 0.45 Logits

DIF of 0.55 logits. Table 4.6 below listed the false positive rate results for a DIF difference of 0.55 logits.

Table 4.6 *Percentage of False Positive Cases across 1,000 Replications with a Difference of 0.55 Logits*

Number of DIF Items	Fit Index	Cut-off Values	Number of Test Takers 5,000	10,000	25,000	50,000
30 Items						
1	MSNQ	0.7, 1.3	85.12	85.23	85.89	87.86
		0.8, 1.2	90.24	90.12	90.77	91.99
	ZSTD	±2.0	3.63	5.08	10.21	19.27
		±3.0	0.26	0.43	1.30	3.86
3	MSNQ	0.7, 1.3	85.77	87.69	91.49	93.73
		0.8, 1.2	90.60	91.76	94.35	95.72
	ZSTD	±2.0	8.61	16.51	37.12	54.47
		±3.0	0.99	3.52	15.28	35.06
6	MSNQ	0.7, 1.3	91.40	96.05	99.41	99.96
		0.8, 1.2	94.22	97.41	99.65	99.98
	ZSTD	±2.0	35.20	62.98	91.73	98.87
		±3.0	9.15	29.50	74.92	93.99
12	MSNQ	0.7, 1.3	99.42	99.98	100.00	100.00
		0.8, 1.2	99.59	99.99	100.00	100.00
	ZSTD	±2.0	93.47	99.20	100.00	100.00
		±3.0	79.69	95.63	99.91	100.00
50 Items						
2	MSNQ	0.7, 1.3	85.19	85.06	85.63	87.24
		0.8, 1.2	90.24	90.16	90.78	91.38
	ZSTD	±2.0	3.58	5.11	9.61	17.42
		±3.0	0.29	0.42	1.09	2.82
5	MSNQ	0.7, 1.3	86.29	87.80	93.60	97.87
		0.8, 1.2	90.99	91.95	95.80	98.55
	ZSTD	±2.0	9.54	18.38	45.11	74.88
		±3.0	1.07	2.98	14.20	41.15
10	MSNQ	0.7, 1.3	94.20	98.26	99.95	100.00
		0.8, 1.2	96.09	98.85	99.99	100.00
	ZSTD	±2.0	48.03	79.14	98.75	99.99
		±3.0	16.46	46.38	91.73	99.71
20	MSNQ	0.7, 1.3	99.88	100.00	100.00	100.00
		0.8, 1.2	99.94	100.00	100.00	100.00
	ZSTD	±2.0	96.64	99.92	100.00	100.00
		±3.0	83.82	99.00	100.00	100.00
70 Items						
3	MSNQ	0.7, 1.3	85.43	85.75	87.71	91.50
		0.8, 1.2	90.37	90.53	91.60	94.37
	ZSTD	±2.0	4.98	8.01	17.86	34.70
		±3.0	0.40	0.83	2.96	9.61
7	MSNQ	0.7, 1.3	87.77	91.06	97.24	99.13
		0.8, 1.2	91.83	94.06	98.21	99.42

14	ZSTD	±2.0	17.10	33.07	69.40	90.18
		±3.0	2.76	8.65	36.32	71.60
	MSNQ	0.7, 1.3	94.53	98.55	99.96	100.00
		0.8, 1.2	96.32	99.02	99.99	100.00
	ZSTD	±2.0	51.00	80.80	98.78	99.99
		±3.0	18.33	49.66	92.68	99.74
28	MSNQ	0.7, 1.3	99.91	100.00	100.00	100.00
		0.8, 1.2	99.94	100.00	100.00	100.00
	ZSTD	±2.0	96.97	99.91	100.00	100.00
		±3.0	86.23	99.09	100.00	100.00

Note: Shaded cells show the minimum/maximum values.

As the magnitude of DIF increased to 0.55 logits, the percentage of false positive rates in MNSQ values did not vary widely. Values ranged from 85.06% to 100.00% with the suggested cut-off values (0.7, 1.3) and from 90.12% to 100.00% for (0.8, 1.2). Again, ZSTD (±2.0) functioned well with a false positive rate close to the Type I error rate of .05 with 4% of the items exhibiting DIF and fewer than 25,000 taking the test. One exception occurred in a 70-item test. The Type I error rate reached 8.01% (or .0801) for 10,000 test takers. For tests with more DIF items and test takers, ZSTD (±3.0) functioned satisfactorily. Specifically, when the proportion of DIF items reached 10%, majority of the testing conditions still achieved a .05 Type I error rate unless the number of test takers reached 25,000.

As seen from the graphs below (Figures 4.11-4.13), when ±2.0 was used to identify item misfit for the OUTFIT ZSTD statistic, the increase in false positive rates was very similar to the graphs for the situation involving a difference of 0.45 logits for DIF. Again, the item fit indices flagged vast majority of the model-fitting items as misfitting (i.e., at least about 90%) when the proportion of 0.55 logits DIF items reached 20% or more.

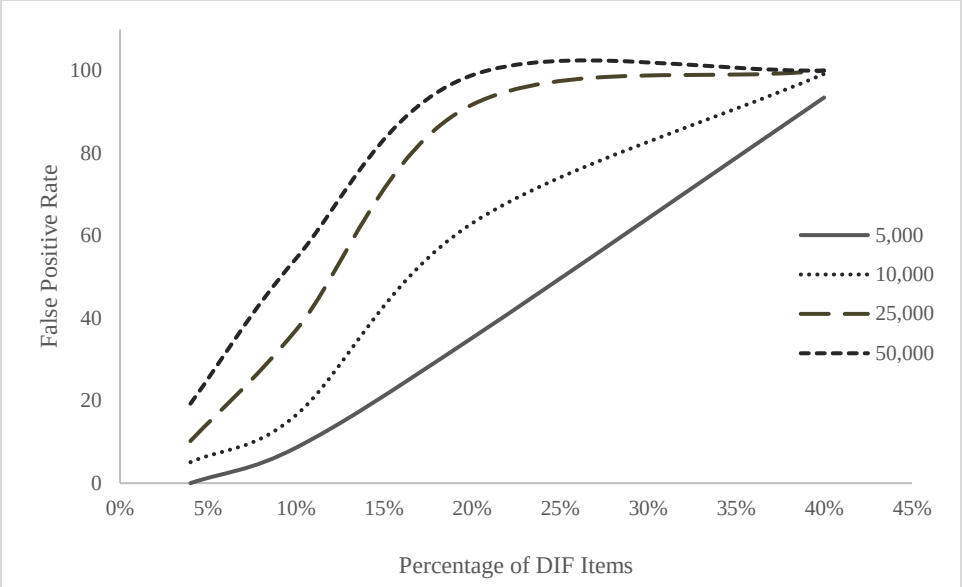

Figure 4.11. False Positive Rates for a 30-item Test with a Difference of 0.55 Logits

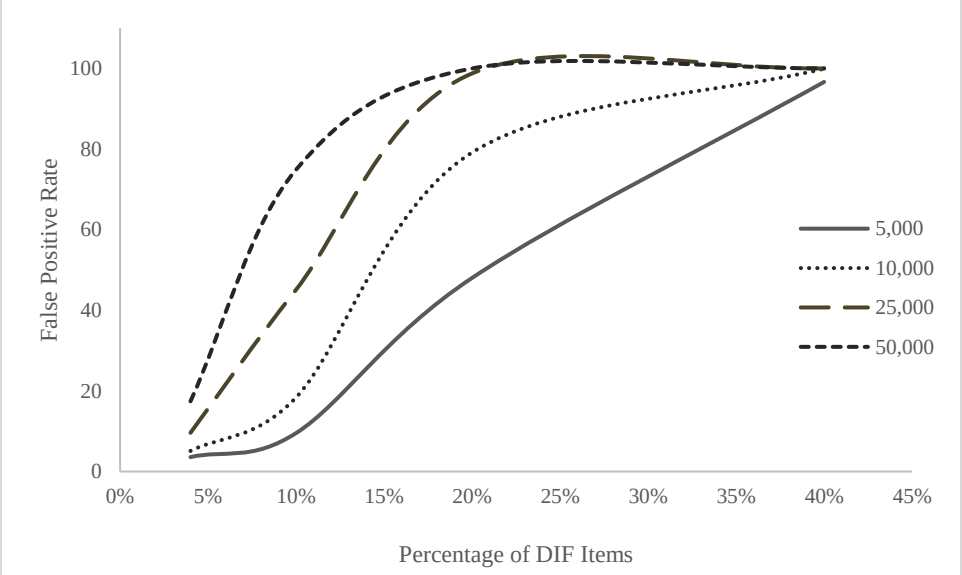

Figure 4.12. False Positive Rates for a 50-item Test with a Difference of 0.55 Logits

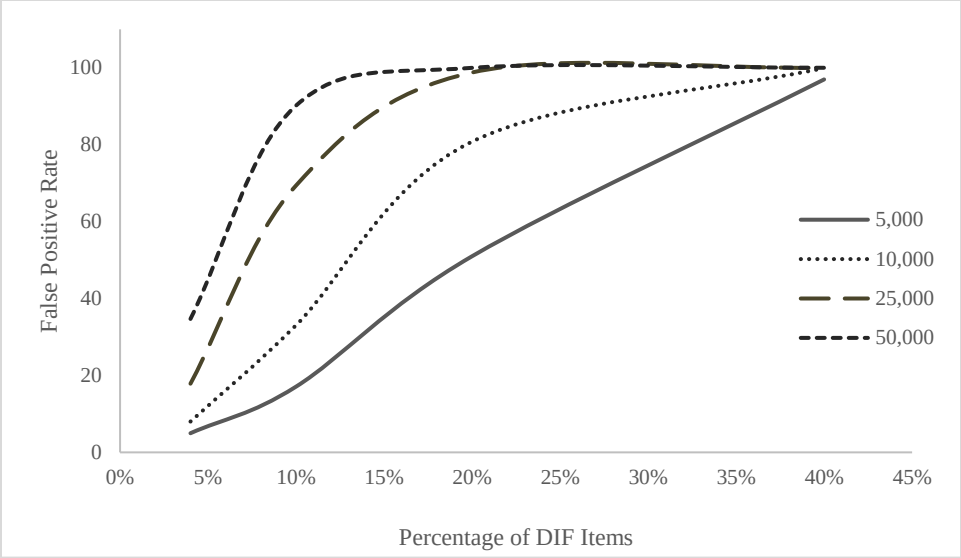

Figure 4.13. False Positive Rates for a 70-item Test with a Difference of 0.55 Logits

***DIF of 0.65 logits*.** False positive rate results for a large DIF difference of 0.65 logits were next presented.

Table 4.7 *Percentage of False Positive Cases across 1,000 Replications with a Difference of 0.65 Logits*

Number of DIF Items	Fit Index	Cut-off Values	Number of Test Takers 5,000	10,000	25,000	50,000
30 Items						
1	MSNQ	0.7, 1.3	85.03	85.50	86.73	89.56
		0.8, 1.2	90.14	90.44	91.33	93.15
	ZSTD	±2.0	4.11	6.27	13.38	25.47
		±3.0	0.32	0.64	1.98	6.26
3	MSNQ	0.7, 1.3	86.46	89.20	91.92	94.39
		0.8, 1.2	91.00	92.87	94.70	96.47
	ZSTD	±2.0	10.84	21.59	43.40	59.71
		±3.0	1.54	5.14	22.42	42.74
6	MSNQ	0.7, 1.3	93.60	97.65	99.78	99.99
		0.8, 1.2	95.68	98.40	99.85	99.99
	ZSTD	±2.0	46.28	74.68	95.98	99.71
		±3.0	15.43	43.81	86.04	97.68
12	MSNQ	0.7, 1.3	99.77	99.98	100.00	100.00
		0.8, 1.2	99.83	99.99	100.00	100.00
	ZSTD	±2.0	96.88	99.66	100.00	100.00
		±3.0	88.97	98.13	100.00	100.00

50 Items						
2	MSNQ	0.7, 1.3	85.24	85.37	86.37	88.85
		0.8, 1.2	90.27	90.31	90.92	92.53
	ZSTD	±2.0	4.03	6.26	12.48	23.59
		±3.0	0.34	0.60	1.65	4.81
5	MSNQ	0.7, 1.3	86.91	89.55	95.60	99.06
		0.8, 1.2	91.44	92.97	97.12	99.44
	ZSTD	±2.0	12.60	25.18	58.27	86.07
		±3.0	1.64	5.24	24.00	59.56
10	MSNQ	0.7, 1.3	96.19	99.45	99.99	100.00
		0.8, 1.2	97.52	99.65	99.99	100.00
	ZSTD	±2.0	62.64	89.84	99.69	100.00
		±3.0	27.56	65.59	97.79	99.99
20	MSNQ	0.7, 1.3	99.98	100.00	100.00	100.00
		0.8, 1.2	99.99	100.00	100.00	100.00
	ZSTD	±2.0	99.15	99.99	100.00	100.00
		±3.0	94.20	99.87	100.00	100.00
70 Items						
3	MSNQ	0.7, 1.3	85.32	86.20	89.30	93.52
		0.8, 1.2	90.32	90.85	92.83	95.79
	ZSTD	±2.0	6.14	10.42	24.63	46.54
		±3.0	0.522	1.34	5.20	16.51
7	MSNQ	0.7, 1.3	89.32	93.47	98.39	99.67
		0.8, 1.2	93.10	95.67	98.91	99.74
	ZSTD	±2.0	23.57	45.37	81.66	95.56
		±3.0	4.81	15.24	54.52	84.96
14	MSNQ	0.7, 1.3	96.73	99.51	99.99	100.00
		0.8, 1.2	97.81	99.69	100.00	100.00
	ZSTD	±2.0	65.56	90.84	99.78	100.00
		±3.0	30.47	68.50	97.93	99.98
28	MSNQ	0.7, 1.3	99.98	100.00	100.00	100.00
		0.8, 1.2	99.99	100.00	100.00	100.00
	ZSTD	±2.0	99.30	99.99	100.00	100.00
		±3.0	95.08	99.90	100.00	100.00

Note: Shaded cells show the minimum/maximum values.

With an item difference of 0.65 logits, the percentage of false positive rates in MNSQ values ranged from 85.03% to 100.00% with the traditional cutoffs 0.7 and 1.3, and from 90.14% to 100.00% using cut-off values of 0.8 and 1.2. In contrast, ZSTD performed satisfactorily, exhibiting a false positive rate close to the Type I error rate of

.05 when a small number of DIF items were present (e.g., 4% and 10% of the whole test) and sample sizes of examinees were smaller than 25,000.

Similarly, the false positive rates skyrocketed as the number of DIF items increased. The percentages of falsely identified non-DIF items jumped to over 90% when the DIF items increased to 20% and 40% of the test items under all test sizes when the MNSQ values were used. This trend began when the difference in DIF was 0.55 logits and was further evident when the DIF difference increased to 0.65 logits. For ZSTD values, when the proportion of DIF items only constituted 4% of the test, ±3.0 could still be used to identify item misfit within the nominal Type I error rate range for the conditions tested here, except for the largest number of test taker 50,000 and the longest test of 70 items (i.e., the false positive rate was 16.51% for this situation).

If ±2.0 was used instead to identify item misfit, only a small number of DIF items (i.e., 4% of the test) and test takers (i.e., 5,000) might be considered for adequate accuracy. Again, as seen from the graphical display below (Figures 4.14-4.16), the false positive rates jumped rapidly as the number of items with large DIF increased. This increase almost reached the maximum when the proportion of DIF items reached 20% of a 30- or 50-item test and 10% for a 70-item test. Basically, the item fit indices flagged all items as misfitting.

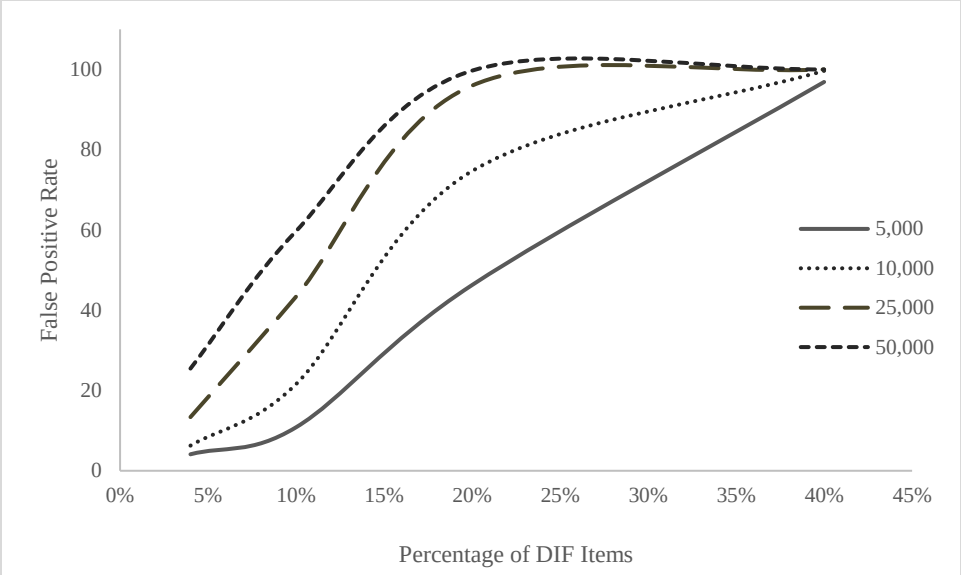

Figure 4.14. False Positive Rates for a 30-item Test with a Difference of 0.65 Logits

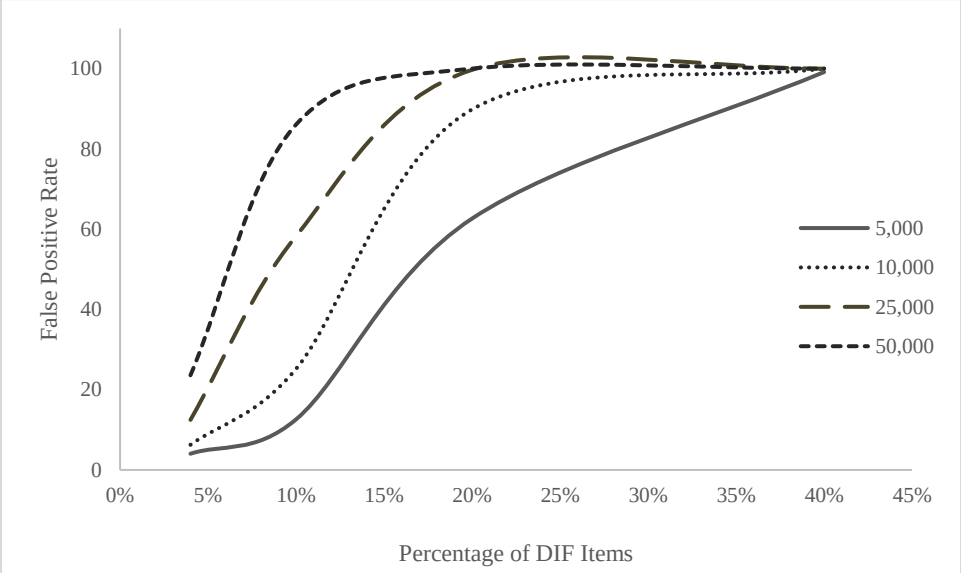

Figure 4.15. False Positive Rates for a 50-item Test with a Difference of 0.65 Logits

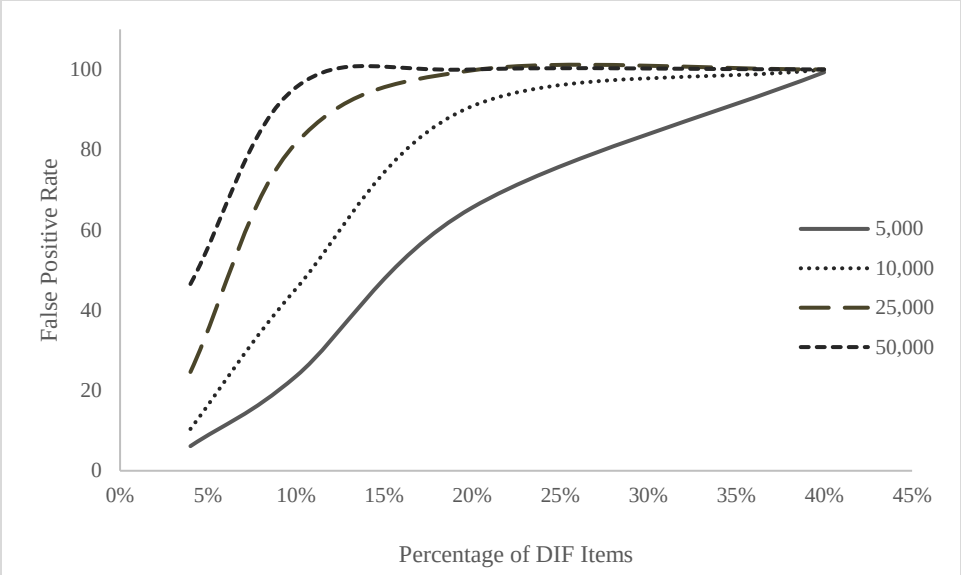

Figure 4.16. False Positive Rates for a 70-item Test with a Difference of 0.65 Logits

In summary, if MNSQ values were used to identify item misfit, the false positive rates were high (i.e., at least 80%) in all testing situations, regardless what cut-off values (i.e., 0.7 and 1.3, 0.8 and 1.2) were adopted. Major contributors to this result may be the large amount of DIF presence and a large number of test takers. Even with a small amount of measurement disturbance, more items may be wrongfully flagged as misfitting. In contrast, if ZSTD values were used, satisfactory results were obtained in some testing situations. Generally, the cut-off values of ±2.0 performed well when the proportion of DIF items constituted no more than 4% of a test with no more than 10,000 test takers for a test of medium length (e.g., 50 items) with moderate item DIF (e.g., 0.35 and 0.45 logits). If the percentage of these biased items reached 10%, the cut-off values of ±3.0 were recommended. For items with large bias (e.g., 0.55 and 0.65 logits), all these cut-off values were still effective as long as the number of test takers did not go beyond 5,000.

CHAPTER 5

DISCUSSION

The purpose of this study was to investigate the accuracy of the mean square (MNSQ) values and the standardized (ZSTD) values of the weighted INFIT and the unweighted OUTFIT total item fit statistics to detect ill-fitting items within the Rasch dichotomous framework. This issue is important to help psychometricians and test developers in selecting and using high quality objective (e.g., multiple-choice) items. As many educational state-wide testing situations use the Rasch model and may be high-stakes for pupils, this investigation is needed.

A simulation study was conducted where conditions of sample size, test length, magnitude and proportion of DIF items were manipulated. Factors were chosen to reflect possible influence from these conditions suggested by previous studies (Awuor, 2008; Miller & Linn, 1988; Miller & Oshima, 1992; Smith et al., 1998) as well as state testing situations (e.g., SC READY, SCPASS, EOCEP).

Specifically, the commonly used Rasch-based fit indices (i.e., the INFIT and the OUTFIT statistics) expressed in MNSQ and ZSTD values were investigated to determine the item misfit. Traditionally, 0.7 and 1.3 are recommended as the cut-off values for MNSQ while ±2.0 for ZSTD values (Wright & Linacre, 1994). In high-stakes tests, more conservative values of 0.8 and 1.2 (Wright & Linacre, 1994) as well as ±3.0 (Iramaneerat et al., 2007) have been suggested.

The study also examined the Type I error rates and the suggested rule-of-thumb critical values under various sample size (e.g., 5,000, 10,000, 25,000, and 50,000) and test length (e.g., 30, 50, and 70 multiple-choice items) conditions in large-scale testing situations. Furthermore, the study evaluated how the obtained values performed to detect simulated measurement disturbances exhibiting through two-group uniform DIF items with a balanced design through between-item fit statistics. The percentage of items that exhibited moderate to large DIF (category "C" per Educational Testing Service's guidelines) ranged from 4%, 10%, 20% and 40% of the test; and items with various degrees of DIF presence ranged from 0.35, 0.45, 0.55 to 0.65 logit units.

Study Findings

Results from the current study affirmed findings from some previous research yet noted unexpected conclusions about the false positive rates (i.e., Type I error rate) of the Rasch-based fit indices.

Research Questions 1 and 2. Results of the simulation were consistent with those provided previous studies (Smith et al., 1998; Wang & Chen, 2005; Wolfe, 2013) as ZSTD values were less influenced by the number of test takers and test items, and that they were more effective than their MNSQ counterparts in detecting item misfit for both INFIT and OUTFIT statistics when items fit the Rasch dichotomous model. The MNSQ values for all INFIT statistic and majority of OUTFIT statistic fell within the 0.9 to 1.1 range, suggesting that the simulation was on target. The optimistic values may be contributed to the large sample sizes studied; as the chi-square values increased with the sample size thus the majority, if not all, items would exhibit acceptance fit to the Rasch model. A large chi-square value may result in under-identification of misfitting items if

the majority or all items 'appear to' fit well when the sample size is large enough. Furthermore, Winsteps® uses a sampling replacement procedure to bootstrap data sets, which might not provide the necessary stochastic structure and thus limit the performance of fit indices.

For ZSTD values, the commonly used values of ±2.0 achieved an error rate lower than .05 for INFIT and OUTFIT total item fit statistics, even though the OUTFIT index had a slightly higher error rate. This means the probability of falsely mis-identifying an item as not fitting the Rasch model was no more than 5% for each of the testing conditions if values of ±2.0 were used as a quality control mechanism. Therefore, both the INFIT and OUTFIT could be recommended for use when the items fit the model. Furthermore, ZSTD values of ±2.0 were suggested over other values like ± 1.0 because other values would result in higher error rates beyond the acceptable level.

In summary, ZSTD values of the INFIT and OUTFIT statistics may work effectively and efficiently to aid test developers in multiple-choice item construction for large-scale assessment programs. Test users may benefit from improvement of high-quality items that would provide accurate information on their test performance through valid score interpretation.

Research Question 3. The broader issue is to identify how these indices performed when errors were present. Since the systematic error (i.e., DIF) was simulated due to programming constraints, the between-item fit statistics were used instead of the total item fit statistics as in the previous two research questions. When the magnitude of DIF was moderate to large (e.g., 0.35 to 0.65 logits), the MNSQ and ZSTD values for the OUTFIT item between-fit statistics performed extremely well when detecting misfitting

items, regardless of sample size, test length, proportion of DIF items present in the test, and the magnitude of DIF. All these indices (i.e., 0.7 and 1.3 as well as 0.8 and 1.2 for MNSQ OUTFIT; ±2.0 and ±3.0 for ZSTD OUTFIT) under all pre-determined testing conditions achieved a hit rate (i.e., true positive rate) of 79% or more, higher than the 70% threshold for acceptable power suggested by González-Romá et al. (2006), indicating a great level of accuracy in result interpretation so as to improve test validity.

Results suggested that the two indices had high power (i.e., ability to correctly identify misfitting cases, true positive rates) in detecting moderate to large uniform DIF items in large-scale testing situations. These indices may be recommended to identify item misfit if only power needs to be taken into consideration. The high power observed here, however, might be due to the influence of sample size on chi-square values, especially when the sample sizes were very large. In statistics, large sample sizes inflate observed power, by making even small deviations observable. Overall, the MNSQ values slightly outperformed their ZSTD counterparts in hit rates, and the results were not significantly different when either value was used. Identification of item misfit using either value would reach at least 90% agreement most of the time (i.e., 99.6% of all estimates). It was recommended that both values may be used.

While high power was evident under these situations, however, it was extremely alarming that the number of well-functioning items being falsely identified as misfitting (i.e., false positive, Type I error) was staggeringly high when MNSQ values were used. Thus, the majority, if not all, test items would be identified as misfitting when the commonly used values (i.e., 0.7 and 1.3; 0.8 and 1.2) were adopted. This would result in unnecessary revision even discard of a majority of well-performing items, leaving test

developers with a much smaller item pool for test development. As this item pool is typically used to create equivalent forms, test developers may encounter the problem of including an inadequate number of unique items (i.e., items that are used in one test form but not the other), thus defeating the purpose of constructing equivalent forms.

If ZSTD values were used, the index could satisfactorily classify DIF items with great accuracy unless the proportion of DIF items was large and/or the number of test takers was very large. Specifically, in these large-scale testing situations, the cut-off values ±2.0 for the ZSTD appeared to be effective in detecting item misfit when only around 4% of items exhibited systematic moderate DIF (e.g., 0.35 and 0.45 logits) with no more than 10,000 test takers for a test of 50 items or fewer. Furthermore, ±3.0 was suggested for using to examine item misfit when the proportion of DIF items reached 10% of the test. For items with large DIF (e.g., 0.55 and 0.65 logits), values of ±2.0 and ±3.0 still worked effectively in the above scenarios when the number of test takers reduced to around 5,000. For a longer test of 70 items, these values were effective with 5,000 test taking students and moderate DIF items. When a large percentage of DIF items (i.e., 40%) was present, either statistic performed well and would identify majority or all items as ill-fitting. These results were somewhat different from the belief that most items will likely be rejected if ZSTD values are used due to the significance of small amount of misfit in large sample situations (Bond & Fox, 2012). Findings from the simulation study suggested that ±2.0 or ±3.0 as a cutoff for the ZSTD may be used in large-scale testing situations unless the proportion of biased items reached 40% of the test.

The use of MNSQ values to detect DIF is more complicated. Basically, the between-item fit indices flagged all items as misfitting in all the large-scale testing situations under study. The hit rates and false negative rates increased with the proportion of flawed items, the number of test-taking individuals, and the magnitude of DIF. These three variables brought more error in comparison to test length conditions. Over-identification of item misfit would result in unneeded revision or even exclusion of majority of the items.

Overall, the ZSTD values outperformed the MNSQ values in false positive rates (i.e., Type I error rates). When the proportion of DIF items were small, and/or the number of test-taking individuals did not reach 50,000, the outperformance could be substantial. This finding extended prior research and provided psychometricians and test developers with information under large-scale testing situations in medium to long tests.

With a large percentage of DIF item presence and/or a large number of test takers, even the ZSTD values (both ±2.0 and ±3.0) resulted in high false positive rates. Therefore, such situations, neither ZSTD nor MNSQ indices could be recommended for use to detect item misfit. Instead, it is recommended that before making a decision about which indices and what values to use for item identification, we need to consider the number of items possibly exhibiting DIF and examine the number of test takers needed. In general, ±2.0 may be recommended to use when the proportion of DIF items are around 4% with 5,000 test takers. For even larger sample size settings, ±3.0 may be suggested. However, caution is recommended because high false positive rates were still observed under in certain situations (e.g., 40% flawed items with large DIF). In such situations, alternative measures may be conducted. For example, sampling procedures

may be considered to obtain reliable item information from a representative group of test takers. Alternatively, pilot studies on items, especially ill-fitting items may be conducted prior to test administration.

The DIF analysis conducted through Winsteps® yielded some unexpected results, with high false positive rates identifying item misfit for OUTFIT ZSTD values in situations of a large amount of measurement error. However, this may be attributed to the procedure used by the Winsteps® software program to examine DIF. In place of the single calibration method that uses the subpopulation residuals to detect item DIF (Smith, R., 2004a), results from Winsteps® between-item fit statistic use a logistic regression analysis. Developed by Swaminathan and Rogers (1990), the logistic regression procedure tests if the logistic regression curves for two groups are the same across ability levels. This framework for conducting a DIF analysis uses a statistical model which is based on the probability of correct response to an item by group membership and a criterion/conditioning variable (e.g., usually the scale or subscale total score). The model uses the item response (0 or 1) as the dependent variable, with three independent variables consisting of the group membership (i.e., reference or focal group), the total score for each respondent, and an interaction between group and total score. Thus, the logistic regression method determines DIF conditioning on the relationship between the item response and the total score, testing the effects of group for uniform DIF or the interaction for non-uniform DIF (Zumbo, 1999).

The logistic regression method was created for use with IRT-based methods, and in particular, the 3PL model. Thus, use of this underlying model is inconsistent with the Rasch measurement framework because the probability of a correct response is not

dependent upon solely the test taker's ability estimate and an item's difficulty estimate. More importantly, use of the logistic regression method to detect DIF is extremely sensitive when sample size is large, thus providing results which are more in line with the use of the total fit statistic (of Winsteps®) and not what is expected if the results were computed using a between-fit statistic (R. Smith, personal communication, July 13, 2020). While findings were unexpected, this information is useful as Winsteps® users should exercise caution if using the between-fit statistic calculated by the program to examine item DIF.

Implications for Practice

A major takeaway message from this study indicated the great impact of measurement disturbances on item performance. When a large amount of error was present, none of the between-item fit indices correctly identified items that were well-fitting. Results showed that many items with error greatly impacted the ability to correctly classify model-fitting items. This suggested that it is imperative for a test to include high quality items, thus making the accurate identification of item misfit more important. In large-scale testing situations, especially when the stakes are high for making crucial decisions on promotion, graduation, or certification, assessments need to use high quality items with solid psychometric properties to safeguard the accuracy of inferences and conclusions drawn from assessment results.

This simulation study showed meaningful implications on item quality to assessment as well as the impact on test validity. Researchers have dedicated long-term efforts in constructing multiple-choice items during the test development process. Proposed and validated by Haladyna and Downing (1989a; 1989b), a detailed description

of 43 principles was provided for general item writing in procedures and content, for stem construction, and for option development that included the correct answer and distractors. It was recommended that these specific research-based principles be followed to guide the construction of effective multiple-choice items (Downing & Haladyna, 1997; Haladyna, 2004).

A study on the consequences of using flawed multiple-choice items in a medical achievement exam by Downing (2005) found that these items were more difficult than their comparable standard questions and resulted in lower passing rates. Another study on continuing medical education (Stagnaro-Green & Downing, 2006) concluded that the items that did not follow the evidence-based item writing guidelines might have introduced construct-irrelevant variance to the assessment and that the variance might have impacted the difficulty of a test item as well as the content independence of a test question. This impact could result in "erroneous test scores and pass-fail decisions" (p. 567). Again, research on high-stakes nursing assessments (Tarrant & Ware, 2008) discovered that high-achieving students were more likely to be penalized by flawed items in comparison to borderline students. This result was consistent with other findings in which flawed questions performed worse and negatively affected student achievement. Furthermore, researchers stated that the impact of such items could be substantial if large numbers of items across multiple assessments exhibited flaws.

In the education domain, researchers (Allalouf & Abramzon, 2008) conducted a study and examined second language assessments. They specifically investigated DIF and concluded that if many items (i.e., 42% in this study) functioned differentially, this could pose a potential threat to validity. Thus, it was suggested to reduce the number of

DIF items while taking into consideration the test's specifications. The current study on power analysis provided some information on the cut-off values to possibly use for various conditions. It is suggested that ±2.0 be used for ZSTD values to identify multiple-choice item misfit when the proportion of DIF items does not exceed 20% due to the high false positive rates. For low-stakes situations like formative assessments, "whose primary goal is assessing the teaching and learning process while it is occurring" (Stout, 2002, p. 506), educators may use the score information with caution and examine individual student's responses to further determine their performance, especially those items with possible bias.

The present effort demonstrated that the amount of error greatly impacted the ability of between-fit item fit statistic to correctly identify well-fitting items. When/If the proportion of DIF items reached 40% of a test, none of the fit indices performed well in item identification within a reasonable error range (i.e., the typical Type I error rate of .05). This serves to caution psychometricians and test developers while constructing multiple-choice items for assessments. Only items with solid psychometric properties render accurate score interpretation and provide solid validity evidence. Correct identification of well-fitting and poor-fitting items marks the initial step in developing an unflawed test question. Over-identification of item misfit would lead to unnecessary revision or exclusion of well-performing items and waste time, money, and efforts dedicated in construction of these items. Thus, the multiple-choice item writing principles are strongly suggested to be adhered to. Furthermore, statistically sound sampling methods may be employed to obtain representative groups of test takers during the item analysis procedure in large-scale testing situations when the number of

examinees is extremely large (i.e., more than 25,000). Pilot studies on all items and a smaller sample of test takers are also recommended to determine item fit prior to the official launch of a test administration.

Limitations and Future Studies

As simulation studies can only manipulate a number of conditions to examine due to various factors (e.g., complexity of a study design, computing power of a statistical/measurement program, programming requirement for a software), findings may be limited to specific situations under investigation. Thus, further research is recommended.

Limitations. As the results suggested, MNSQ values were problematic in identifying total item fit under the Rasch dichotomous model. To remedy this, a study conducted by Wu and Adams (2013) examined the properties of residual-based fit indices in Rasch measurement models and analytically illustrated that these statistics "provide a measure of the relative slope of empirical item characteristic curves" (p. 339) for dichotomous responses. It concluded that employing a sample size formula $1 \pm 2\sqrt{\frac{2}{N}}$ to approximate (p. 352) could provide an acceptable range for the MNSQ values. Simulation studies may be conducted to further determine how effective this criterion is before this rule of thumb becomes common practice. In the meantime, ZSTD values are recommended and the customary ±2.0 may be used in large-scale testing situations. Furthermore, based on previous studies (Seol, 2016; Su et al., 2007), it is suggested that effect size measures are needed to show the magnitude of fit, thus CIs for the INFIT and OUTFIT mean square errors may be used and reported (Bandalos & Leite, 2013).

Fit is well known that including construct-irrelevant variables within a test is detrimental to the validity of the associated test scores. All psychological measurements include measurement error to some extent. These errors can be broadly categorized as systematic error and random error (Bond & Fox, 2012). Systematic errors impact a test taker's raw score in a consistent manner due to some particular characteristic of the person or the test (i.e., gender, ethnicity, biased items, testing format that is familiar to a particular group of test takers). The characteristic is independent of the construct being measured (Bond & Fox, 2012). Random errors, on the contrary, affect a test taker's observed score due to chance happenings (Bond & Fox, 2012). For instance, a test taker may be more familiar with some content of the testing materials covered, distraction in the testing room, error during test administration, etc. These errors affect the scores in positive or negative direction (Bond & Fox, 2012). In the present study, only systematic error was simulated in the data sets due to programming constraints. A study conducted by Smith (1994) investigated the power of item fit indices in detecting misfit. Comparing the total item fit and the between item fit indices, it concluded that total fit statistic was less sensitive to systematic error and lacked power in detecting systematic measurement disturbances like item bias. Furthermore, both the weighted and the unweighted between-fit statistics performed similarly enough so that one version of these was sufficient to use. Therefore, between-fit statistic was used to examine performance of item fit indices due to their high power to detect systematic error in comparison to total fit statistic; thus, the performance of total fit statistics needs further investigation. Future studies would like to generate response data with random error (i.e., guessing,

carelessness) in order to provide a better glimpse of the power and Type I error rates of total item fit indices.

While it is encouraging to obtain an extremely high power of correctly identifying biased items with moderate to strong DIF, it is more meaningful to understand how the magnitude of bias influences the performance of item fit statistics. The present study has attempted to examine items with moderate to large DIF; however, in reality, items with substantial DIF are usually revised and removed during item review sessions. Major assessment programs, especially those that have been implemented multiple years, are usually well-designed. Furthermore, it is rare to encounter tests whose items have the same amount of bias or uniform DIF in groups. Studies on more complex situations that more realistically mirror real-world situations can be designed. For instance, items with small DIF are considered negligible and it is common to have such items in assessments. While the number of test takers is extremely large, or with a high percentage of DIF items (i.e., 20% or more), will the amount of error reach an impactful level to influence the performance of item fit indices? Other situations that warrant careful examination include: items with non-uniform DIF like ordinal DIF (i.e., item difficulty is different for the same group of examinees across all levels of ability; Walker, 2011) and disordinal DIF (i.e., item difficulty is different for a group of examinees at different ability levels; Walker, 2011); groups of unequal number of test takers (i.e., unbalance design like ethnicity) and multiple groups of test takers categorized by group membership; a mixture of items with small to moderate DIF and moderate to large DIF; bias items exhibit random DIF as well as systematic DIF, etc. Nevertheless, this may require computers and calibration programs to have high computing capacity and to be programming

uncomplicated in generating the response data that meet specific study needs. For low-stakes assessments, educators may use score information with great caution. Careful examination of student responses, especially those that are highly unexpected may be necessary.

While results on the power analysis and the Type I error rates from the current study provided an excellent springboard to future research, the Smith study (1996) investigating the power of between-item fit statistic used unbalanced design with low to high biased items. This study also concluded that the number of unbiased items being misidentified (i.e., false positive cases, Type I error rates) increased as the number of biased items increased. However, the Smith study (1996) found that the bias magnitude needed to be larger than 0.5 logits in order to be detected more than 50% of the time under the situation of 100 samples in the focal group and 1,000 in the reference group. Also different from the current study, this unbalanced design study (Smith, 1996) indicated that the power to detect item bias decreased as the number of biased items increased and as the number of falsely identified unbiased items (i.e., false positive cases, Type I error rates) increased with smaller numbers of biased items. Future work on the power and Type I error rates needs to be furthered.

Future research. Even though the current study could not identify specific sample sizes for various testing conditions to obtain reliable cut-off values for the item fit statistics, it has again stressed that sample size has an effect on chi-square type of statistic. It is usually stated that these statistics do not function well when sample size is large (Linacre, 2019a). However, it has not been made explicit how large is considered large. No specific guidelines have been proposed. Therefore, it is imperative that such

studies be conducted to guide researchers in choosing an appropriate number to pivot their study.

In addition, the study generated data that fit the Rasch dichotomous model. In other words, correctly specified models were used. It is as important to examine mis-specified models to see how these indices behave. For instance, the DeMars' study (2017) simulated several misfitting items (i.e., items with a different slope, items with a guessing parameter) and concluded that it appeared to be reasonable for judging the magnitude of misfit in the sample if the fit was statistically significant. Inclusion of poorly fitting items better mirrors real-world testing situations that both well-functioning and misfitting items may be present in a test. This warrants further research in future studies. Additionally, state-wide assessment programs utilize test items other than the multiple-choice format (e.g., short-answer questions in SC Ready [SC Department of Education, 2016]; constructed-response items in SCPASS [SC Department of Education, 2008]; essay writing in EOCEP [SC Department of Education, 2006]; etc.). These types of response data (i.e., ordinal scale like a rating scale, answers that warrant a partial credit) could also be included for more complex tests in follow-up studies.

It is noted that real world data can be complex (Bandalos & Leite, 2013) and empirical data are more likely to reflect the data complexity encountered in real world situations. Thus, future studies may use empirical data to validate the study results. For instance, music assessment data from the 2016-2017 test administration of the South Carolina Arts Assessment Program (SCAAP; SCDE, 2003) may be used to provide person and item estimates to generate response patterns. Instead of the standard normal distribution for the person ability estimates and the uniform distribution to estimate item

difficulty parameter, specific estimates on person ability and item difficulty may be obtained, utilized, and simulated to generate responses. The total number of items from this empirical data includes 45 multiple-choice items and one test form, with a range of difficulty levels (i.e., *p*-values) from 0.20 to 0.87. Studies from this data may warrant interesting findings.

As it was found that the Winsteps® between-item fit statistics used the logistic regression framework instead of the true between-item fit statistic, this finding should be examined further. For example, a comparison of the regression-based Winsteps® results with procedures more aligned with the Rasch framework, such as the separate calibration results based upon standard error (i.e., the *t*-test method using the differences between two subpopulations on the same item for a pairwise comparison[4]; Wright & Stone, 1979) as well as calculations that use between-item fit statistic (see Formulas [10]-[13], Smith, 1994). In addition, future studies could attempt to replicate the findings using different sample sizes and levels of DIF among groups.

Further, multiple programs (i.e., WinGen, Winsteps®) and software (i.e., R) were used to generate data. Different programs and software may cause inconsistency in data simulation and impact study results. Future studies may use one software program (e.g., R for data generation and result analysis) to determine if the results found here can be replicated. Finally, the bootstrapping approach used in Winsteps® warrants in-depth examination as the sampling method can impact the performance of item fit statistics, and hence, the interpretation of the results.

---

[4] The formula is $t = \frac{d_{i1} - d_{i2}}{(s_{i1}^2 + s_{i2}^2)^{1/2}}$, [14]
where d_{i1} and d_{i2} are the difficulty of item *i* in the calibration based on each subpopulation, s_{i1} and s_{i2} are the standard error of estimate for d_{i1} and d_{i2} (Smith, 2004b).

Summary and Significance of the Study

The present study simulated dichotomous response data to examine the performance of four total item fit statistics: the weighted INFIT mean square, the weighted INFIT standardized value, the unweighted OUTFIT mean square, and the unweighted OUTFIT standardized value. In large-scale testing situations where the number of test takers easily reach thousands and tens of thousands, the mean square values over-identify unbiased items as the chi-square statistic is extremely sensitive to sample size. Simulation results suggested that the traditionally used ±2.0 values may be used for the standardized values to correctly identify item misfit at within the acceptable Type I error rate of .05. Investigation on the power and Type I error rate of between-item fit statistics to identify simulated systematic measurement disturbances still needs further study.

This study filled a gap in the Rasch literature in a number of ways. First, the research approximated typical testing practices in the field by investigating the impact of larger sample sizes and longer tests on information used to make decisions about item fit. Previous studies in educational research did not have the computation capacity to mirror large-scale assessment situations as described in this study. As many tests use the rule-of-thumb values, it is crucial that these indices can detect items which do not fit the model (i.e., measurement disturbances) with sufficient power. It is hoped that findings provide useful information to psychometricians and large-scale test developers in selecting objective test items that are well-fitting under a range of conditions (i.e., numbers of test taker, test length, proportion and magnitude of DIF items) that are similar to those encountered in practice. Furthermore, the high importance of item quality needs

for psychometrician and test developers to take heed because measurement error greatly impacts item performance. Even though it would be unexpected to have 40% of test items with DIF, the bounds of the indices were tested with simulation. While it is very likely to have 4% of items with DIF, it is hoped that concrete results on the power and Type I error rates of the item fit indices will provide test developers with useful information to guide practice, especially the use of true item fit statistics aligned with the Rasch framework.

REFERENCES

Adams, R. J., & Wilson, M. (1996, April). *Multilevel modeling of complex item responses in multiple dimensions: Why bother?* Paper presented at the annual meeting of AERA, New York, NY.

Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Alasuutari, P., Bickman, L., & Brannen, J. (2009). *The SAGE handbook of social research methods*. London, UK: SAGE.

Allalouf, A., & Abramzon, A. (2008). Constructing better second language assessments based on Differential Item Functioning analysis. *Language Assessment Quarterly, 5*(2), 120-141. doi: 10.1080/15434300801934710

American College Testing (1959-2020) [Measurement instrument]. Iowa City, IA: ACT Inc.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

An, X., & Yung, Y.-F. (2014). *Item response theory: What it is and how you can use the IRT procedure to apply it* (Paper SAS364-2014). Cary, NC: SAS Institute. Retrieved on July 12, 2019 from https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123-140.

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Andrich, D., & Marais, I. (2018). Controlling bias in both constructed response and multiple-choice items when analyzed with the dichotomous Rasch model. *Journal of Educational Measurement*, *55*(2), 281-307.

Andrich, D., Marais, I., & Humphry, S. M. (2016). Controlling guessing bias in the dichotomous Rasch model applied to a large-scale, vertically scaled testing program. *Educational and Psychological Measurement*, *76*(3), 412-435. doi: 10.1177/0013164415594202

Andrich, D., Styles, I., Mercer, A., & Puddey, I. B. (2017). On the validity of repeated assessments in the UMAT, a high-stakes admissions test. *Advances in Health Sciences Education*, *22*, 1245-1262.

Awuor, R. A. (2008). Effect of unequal sample sizes on the power of DIF detection: An IRT-based Monte Carlo study with SIBTEST and Mantel-Haenszel procedures (Doctoral dissertation). Retrieved on June 27, 2019 from https://vtechworks.lib.vt.edu/bitstream/handle/10919/28321/RAA_ETD.pdf?sequence=1&isAllowed=y

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions*, *21*(3), 1105-1106. Retrieved on January 6, 2020 from https://www.rasch.org/rmt/rmt213b.htm

Baghi, H. (1990). *The use of Rasch model fit statistics in selecting items for the Maryland Functional Testing Program* (Report). Baltimore, MD: Program Assessment, Evaluation, and Instructional Support System, Maryland State Department of Education. Retrieved on October 8, 2019 from https://files.eric.ed.gov/fulltext/ED324363.pdf

Bandalos, D. L., & Leite, W. (2013). Use of Monte Carlo studies in Structural Equation Modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A second course* (2nd ed., pp. 625-666). Charlotte, NC: Information Age Publishing.

Baylor, C., Hula, W., Donovan, N. J., Doyle, P. J., Kendall, D., & Yorkston, K. (2011). An introduction to Item Response Theory and Rasch models for speech-language pathologists. *American Journal of Speech-Language Pathology*, *20*, 243-259.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley Publishing Company, Inc.

Bond, T., & Fox, C. (2012). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110.

Boone, W. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education*, *15*(4), 1-7. doi: 10.1187/cbe.16-04-0148

Bowe, B., & Cronin, J. (2005). *Aligning the NWEA RIT scale with the Maryland school assessment (MSA)* (Report). Lake Oswego, OR: Northwest Evaluation Association. Retrieved on January 18, 2019 from https://files.eric.ed.gov/fulltext/ED491231.pdf

Burton, R. F. (2006). Sampling knowledge and understanding: How long should a test be? In *Assessment & Evaluation in Higher Education*, *31*(5), 569-582.

Camilli, G., & Shephard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: SAGE Publications, Inc.

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

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

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

Dappen, L., Isernhagen, J., & Anderson, S. (2008). A statewide writing assessment model: Student proficiency and future implications. *Assessing Writing*, *13*, 45-60.

DeMars, C. E. (2017). Infit and outfit: Interpreting statistical significance and magnitude of misfit in conjunction. *Journal of Applied Measurement*, *18*(2), 163-177.

DePascale, C. A. (2003, April). *Putting large-scale assessment in perspective: The ideal role of large-scale assessment in a comprehensive assessment system*. Paper presented at the annual meeting of AERA, Chicago, IL.

DeVellis, R. F. (2006). Classical Test Theory. *Medical Care*, *44*(11), S50-S59. Retrieved on June 28, 2019 from http://www.jstor.org/stable/41219505

Differential Item Functioning (n.d.). In *Wikipedia*. Retrieved on July 1, 2020 from https://en.wikipedia.org/wiki/Differential_item_functioning

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items n achievement examinations in medical education. *Advances in Health Sciences Education*, *10*, 133-143.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, *1*, 61-82.

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annuals of Statistics*, *7*, 1-26.

End-of-Course Examination Program (2006-2020) [Measurement instrument]. Columbia, SC: The SC Department of Education (SCDE).

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

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

Engelhard, G., Kobrin, J. L., & Wind, S. A. (2014). Exploring differential subgroup functioning on SAT writing items: What happens when English is not a test

taker's best language? *International Journal of Testing, 14*, 339-359. doi: 10.1080/15305058.2014.931281

Engelhard, G., Wind, S. A., Kobrin, J. L., & Chajewski, M. (2013). *Differential item and person functioning in large-scale writing assessments within the context of the SAT* (Report 2013-6). New York, NY: College Board. Retrieved on January 18, 2019 from https://files.eric.ed.gov/fulltext/ED562620.pdf

Erhart, M., Hagquist, C., Auquier, P., Rajmil, L., Power, M., Ravens-Sieberer, U., & the European KIDSCREEN Group (2009). A comparison of Rasch item-fit and Cronbach's alpha item reduction analysis for the development of a Quality of Life scale for children and adolescents. *Child: Care, Health and Development*, *36*(4), 473-484.

Every Student Succeeds Act. (2015). *An Act to reauthorize the Elementary and Secondary Education Act of 1965 to ensure that every child achieves* (Pub. No. 114-95). Washington, DC: U.S. Government Printing Office.

Every Student Succeeds Act (n.d.). In *Wikipedia*. Retrieved on January 4, 2020 from https://en.wikipedia.org/wiki/Every_Student_Succeeds_Act

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.

Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society A*, *222*(594-604), 309-368. doi: 10.1098/rsta.1922.0009

Fleckenstein, J., Leucht, M., Pant, H. A., & Köller, O. (2016). Proficient beyond borders: Assessing non-native speakers in a native speakers' framework. *Large-scale*

Assessments in Education, *4*(19), 1-19. Retrieved on January 18, 2019 from https://link.springer.com/content/pdf/10.1186%2Fs40536-016-0034-2.pdf

Florin, R. (1999). *Paired comparisons: A method for ranking physician work*. Unpublished working paper.

Global Rasch fit statistic. (2007). *Rasch Measurement Transactions*, *21*(2), 1103. Retrieved on June 28, 2019 from https://www.rasch.org/rmt/rmt212e.htm

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research*, *41*(1), 29-53.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education*, *17*(2), 145-220.

Grinnell, R. (2018). Objective test. *PsychCentral*. Retrieved on May 14, 2019 from https://psychcentral.com/encyclopedia/objective-test/

Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *33*(2), 205-233.

Hagquist, C., & Hellström, L. (2014). The psychometric properties of the early development instrument: A Rasch analysis based on Swedish pilot data. *Social Indicators Research*, *117*(1), 301-317. doi: 10.1007/s11205-013-0344-5

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 37-50.

Haladyna, T. M., & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*, 51-78.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38-47.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.

Harsch, C., & Rupp, A. A. (2011). Designing and scaling level-specific writing tasks in alignment with the CEFR: A test-centered approach. *Language Assessment Quarterly*, *8*(1), 1-33. doi: 10.1080/15434303.2010.535575

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Modeling booklet effects for nonequivalent group designs in large-scale assessment. *Educational and Psychological Measurement*, *75*(4), 568-584.

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., & Epstein, R. (2005). *Bootstrap methods and permutation tests* (2nd ed.). Retrieved on May 23, 2019 from https://hsta559s12.pbworks.com/w/file/fetch/50747070/Hesterberg.Bootstrappng.chpt18.pdf

High School Assessment Program (2009-2014) [Measurement instrument]. Columbia, SC: The SC Department of Education (SCDE).

Hilmer, C. E., & Holt, M. T. (2000, July). *A comparison of resampling techniques when parameters are on a boundary: The bootstrap, subsample bootstrap, and*

subsample jackknife. Paper presented at the annual meeting of the American Agricultural Economics Association, Tampa, FL.

History of Military Testing (n.d.). In *Armed Services Vocational Aptitude Battery*. Retrieved on October 29, 2019 from https://www.officialasvab.com/history_rec.htm

History of Standardized Testing in the United States (n.d.). In *National Education Association*. Retrieved on October 10, 2019 from http://www.nea.org/home/66139.htm

Hodge, K., J., & Morgan, G. (2017). Stability of INFIT and OUTFIT compared to simulated estimates in applied setting. *Journal of Applied Measurement*, *18*(4), 383-392.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysis item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *17*(6), 497-509. doi: 10.1080/13803611

Holland, P. W., & Thayer, D. T. (1985). *An alternate definition of the ETS delta scale of item difficulty* (Report No. 85-64). Princeton, NJ: Education Testing Service. Retrieved on January 7, 2020 from https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2330-8516.1985.tb00128.x

Imperial Examination (n.d.). In *Wikipedia*. Retrieved on October 10, 2019 from https://zh.m.wikipedia.org/zh/%E4%B8%AD%E5%9B%BD%E7%A7%91%E4%B8%BE%E5%88%B6%E5%BA%A6

Imperial Examination in China (n.d.). In *Wikipedia*. Retrieved on October 10, 2019 from https://zh.m.wikipedia.org/zh/中国科举制度

Iramaneerat, C., Smith, E. V., & Smith, R. M. (2007). An introduction to Rasch measurement. In J. W. Osborne (Ed.). *Best practices in quantitative methods* (pp. 50-70). Thousand Oaks, CA: Sage.

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

Kang, S. J., Sim, K. H., Song, B. R., Park, J.-E., Chang, S. J., Park, C., & Lee, M. S. (2018). Validation of the health literacy scale for diabetes as a criterion-referenced test with standard setting procedures. *Patient Education and Counseling, 101*, 1468-1476.

Klassen, A. F., Grant, C., Barr, R., Brill, H., Kraus de Camargo, O., Ronen, G. M., Samaan, M. C., Mondal, T., Cano, S. J., Schlatman, A., Tsangaris, E., Athale, U., Wickert, N., & Gorter, J. W. (2014). Development and validation of a generic scale for use in transition programmes to measure self-management skills in adolescents with chronic health conditions: The TRANSITION-Q. *Child: Care, Health and Development*, *41*(4), 547-558. doi: 10.1111/cch.12207

Koo, J., Becker, B. J., & Kim, Y.-S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing, 31*(1), 89-109.

Kubinger, K. D., Hohensinn, C., Hofer, S., Khorramdel, L., Frebort, M., Holocher-Ertl, S., Reif, M., & Sonnleitner, P. (2011). Designing the test booklets for Rasch model calibration in a large-scale assessment with reference to numerous

moderator variables and several ability dimensions. *Educational Research and Evaluation: An International Journal on Theory and Practice, 17*(6), 483-495. doi: 10.1080/13803611.2011.632666

Landau, J. K. (1998). *Statewide assessment: Policy issues, questions, and strategies* (PEER Policy Paper). Boston, MA: The Federation for Children with Special Needs. Retrieved on July 30, 2019 from https://files.eric.ed.gov/fulltext/ED420934.pdf

LeBeau, B., & McVay, A. (2017). *Validity of the three parameter item response theory model for field test data* (ITP Research Series). Iowa City, IA: Iowa Testing Programs, the University of Iowa. Retrieved on November 4, 2019 from https://itp.education.uiowa.edu/ia/documents/Validity-of-the-Three-Parameter-Item-Response-Theory-Model-for-Field-Test.pdf

Linacre, J. M. (n.d.) *Estimation bias correction – warnings*. Retrieved on January 14, 2020 from https://www.winsteps.com/winman/ebiascorrection.htm

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

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. In S. Chae, U. Kang, E. Jeon, & J. M. Linacre (Eds), *Development of computerized middle school achievement test* (pp. 1-58). Chicago, IL: MESA.

Linacre, J. M. (2002a). Residuals and rating scales. *Rasch Measurement Transactions, 16*(1), 866. Retrieved on June 30, 2020 from https://www.rasch.org/rmt/rmt161.pdf

Linacre, J. M. (2002b). What do Infit and Outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. Retrieved on June 19, 2019 from https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2019a). *Fit diagnosis: Infit Outfit mean-square standardized*. Retrieved on June 28, 2019 from https://www.winsteps.com/winman/misfitdiagnosis.htm

Linacre, J. M. (2019b). *Reliability and separation of measures*. Retrieved on July 1, 2020 from https://www.winsteps.com/winman/reliability.htm#:~:text=Rasch%20Reliability%20is%20higher%20than%203%2DPL%20IRT%20Reliability.&text=Item%20reliability%3A%20Low%20reliability%20means,the%20same%2C%20not%20reliably%20different.

Linacre, J. M. (2019c). Winsteps® (Version 4.4.6) [Computer Software]. Beaverton, OR: Winsteps.com. Retrieved on January 1, 2019. Available from https://www.winsteps.com/

Linacre, J. M. (2019d). *Simulate file specifications*. Retrieved on July 22, 2020 from https://www.winsteps.com/winman/simulated.htm

Linacre, J. M., & Wright, B. D. (1999). *A users' guide to WINSTEPS*. Chicago, IL: MESA Press.

Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. *Applied Measurement in Education*, *22*, 164-184.

Lord, F. M. (1964). *A strong true-score theory, with applications*. Princeton, NJ: Educational Testing Service and Princeton University.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Welsey Publishing Company, Inc.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McQueen, J., & Congdon, P. J. (1997, March). *Rater severity in large-scale assessment: Is it invariant*? Paper presented at the annual meeting of AERA, Chicago, IL.

Mead, R. (1976, April). *Assessing the fit of data to the Rasch model*. Paper presented at the annual meeting of the AERA, San Francisco, CA.

Measures of Academic Progress (2013-2020) [Measurement instrument]. Portland, OR: Northwest Evaluation Association (NWEA).

Mesic, V., & Muratovic, H. (2011). Identifying predictors of physics item difficulty: A linear regression approach. *Physical Review Special Topics Physics Education Research, 7*(1), 1-15. doi: 10.1103/PhysRevSTPER.7.010110

Messick, S. (1996). Validity of performance assessment. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington, DC: National Center for Education Statistics.

Miller, D. (2004). A Macro for reading multiple text files (Paper SUGI057-29). Cary, NC: SAS Institute. Retrieved on June 20, 2019 from https://support.sas.com/resources/papers/proceedings/proceedings/sugi29/057-29.pdf

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

Miller, M. D., & Oshima, T. C. (1992). Effect of sample size, number of biased items, and magnitude of bias on a two-stage item bias estimation method. *Applied Psychological Measurement, 16*(4), 381-388.

Monseur, C., & Adams, R. J. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement, 10*(3), 320-334.

National Assessment of Educational Progress (1969-2020) [Measurement instrument]. Washington, D. C.: National Center for Educational Statistics (NCES).

National Center for Education Statistics (NCES). (2017). *Public high school graduates, by region, state, and jurisdiction: Selected years, 1980-81 through 2026-27* [Data file and code book]. Retrieved on July 26, 2019 from https://nces.ed.gov/programs/digest/d16/tables/dt16_219.20.asp

No Child Left Behind Act of 2001. (2002). *An Act to close the achievement gap with accountability, flexibility, and choice, so that no child is left behind* (Pub. No. 107-110). Washington, DC: U.S. Government Printing Office.

Oliveri, M. E., Ercikan, K., Lyons-Thomas, J., & Holtzman, S. (2016). Analyzing fairness among linguistic minority populations using a latent class differential item functioning approach. *Applied Measurement in Education, 29*(1), 17-29.

Olsen, J. B., Maynes, D. D., Slawson, D., & Ho, K. (1989). Comparisons of paper-administered, computer-administered and computerized adaptive achievement tests. *Journal of Educational Computing Research, 5*(3), 311-326.

Palmetto Assessment of State Standards (2008-2020) [Measurement instrument]. Columbia, SC: The SC Department of Education (SCDE).

Papadopoulos, T. C., Spanoudis, G., & Kendeou, P. (2009). The dimensionality of phonological abilities in Greek. *Reading Research Quarterly*, *44*(2), 127-143.

Parhizgar, S. (2012). Testing and technology: Past, present and future. *Theory and Practice in Language Studies*, *2*(1), 174-178.

Pelton, T. W. (2002). *The accuracy of unidimensional measurement models in the presence of deviations from the underlying assumptions*. (Doctoral dissertation). Retrieved on November 4, 2019 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.474.8405&rep=rep1&type=pdf

Peoples, S. M. (2012). *The Nature of Science Instrument-Elementary (NOSI-E): Using Rasch principles to develop a theoretically grounded scale to measure elementary student understanding of the nature of science* [Measurement instrument]. Ann Arbor, MI: ProQuest LLC.

Peoples, S., O'Dwyer, L., Shields, K., & Wang, Y. (2013). With hiccups and bumps: The development of a Rasch-based instrument to measure elementary students' understanding of the nature of science. *Journal of Applied Measurement*, *14*(1), 57-78.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237-255.

Pilatti, A., Godoy, J. C., Lozano, Ó., & Brussino, S. (2015). Psychometric properties of the Alcohol Expectancy Scale in Argentinean Adolescents applying the rating scale analysis. *Journal of Child & Adolescent Substance Abuse*, *24*, 264-273.

Pilatti, A., Lozano, O. M., & Cyders, M. A. (2015). Psychometric properties of the Spanish version of the UPPS-P Impulsive Behavior Scale: A Rasch rating scale analysis and confirmatory factor analysis. *Psychological Assessment, 27*(4), e10-e21.

Pilatti, A., Read, J. P., & Caneto, F. (2016). Validation of the Spanish version of the Young Adult Alcohol Consequences Questionnaire (S-YAACQ). *Psychological Assessment, 28*(5), e49-e61.

Pilatti, A., Read, J. P., Vera, B., Caneto, F., Garimaldi, J. A., & Kahler, C. W. (2014). The Spanish version of the Brief Young Adult Alcohol Consequences Questionnaire (B-YAACQ): A Rasch model analysis. *Addictive Behaviors, 39*, 842-847.

Pintner, R. (1921). The Army Mental Tests. *The Journal of Educational Psychology, 12*(7), 418-420.

Pompey, K., Jiang, N., Burgess, Y., & Lewis, A. (2019). *A step-by-step application of DIF analysis using a two-stage procedure: The Rasch trees and Mantel-Haenszel methods*. Manuscript in preparation.

Popham, W. J. (2001). *The truth about testing: An educator's call to action*. Alexandria, Virginia: Association for Supervision and Curriculum Development.

Popp, S. E. O., Ryan, J. M., Thompson, M. S., & Behrens, J. T., (2003, April). *The effect of benchmark selection on the assessed quality of writing*. Paper presented at the annual meeting of AERA, Chicago, IL.

Programme for International Student Assessment (2000-2018) [Measurement instrument]. Paris, France: Organization for Economic Cooperation and Development (OECD).

Quan, B., Park, T. A., Sandahl, G., & Wolfe, J. H. (1984). *Microcomputer network for Computerized Adaptive Testing (CAT)* (Report NPRDC TR 84-33). San Diego, CA: Navy Personnel Research and Development Center.

R Development Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (1974). An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods & Instrumentation*, *6*(2), 208-212.

Ritesh, R. (2016). Analysis of performance of indices for Indian mutual funds. In K. Chen & A. Ravindran (Eds.), *Forging connections between computational mathematics and computational geometry*. (pp. 221-233). Basel, Switzerland: Springer International Publishing.

Rusticus, S. A., & Lovato, C. Y. (2014). Impact of sample size and variability on the power and Type I error rates of equivalence tests: A simulation study. *Practical Assessment, Research & Evaluation*, *19*(11), 1-10. Retrieved on July 22, 2019 from https://pareonline.net/getvn.asp?v=19&n=11

Sachse, K. A., Roppelt, A., & Haag, N. (2016). A comparison of linking methods for estimating national trends in international comparative large-scale assessments in

the presence of cross-nation DIF. *Journal of Educational Measurement*, *53*(2), 152-171.

Sälzer, C., & Heine, J. H. (2016). Students' skipping behavior on truancy items and (school) subjects and its relation to test performance in PISA 2012. *International Journal of Educational Development*, *46*, 103-113.

SAS ® (Version 9.4) [Computer software]. (2017). Cary, NC: SAS Institute Inc.

Scarpati, S. E., Wells, C. S., Lewis, C., & Jirka, S. (2011). Accommodations and item-level analyses using mixture differential item functioning models. *The Journal of Special Education*, *45*(1), 54-62.

Scholastic Assessment Test (1993-2020) [Measurement instrument]. New York, NY: College Board.

Schumacker, R. E. (1999). Many-facet Rasch analysis with crossed, nested, and mixed designs. *Journal of Outcome Measurement*, *3*, 323-338.

Schumacker, R. E. (2004). Rasch measurement: The dichotomous model. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 226-257). Maple Grove, MN: JAM Press.

Schumacker, R. E., Mount, R., & Marcoulides, G. A. (2005, April). *Detecting measurement disturbance effects: The graphical display of item characteristics*. Paper presented at the annual meeting of AERA, Montréal, QC.

Seol, H. (2016). Using the bootstrap method to evaluate the critical range of misfit for polytomous Rasch fit statistics. *Psychological Reports*, *18*(3), 937-956.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe,M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*(33), 1-11. doi: 10.1186/1471-2288-8-33

Smith, E. V. (2004). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 93-122). Maple Grove, MN: JAM Press.

Smith, R. M. (1982). *Detecting measurement disturbances with the Rasch model* (Doctoral dissertation). Retrieved on January 5, 2020 from https://search-proquest-com.pallas2.tcl.sc.edu/pqdtglobal/docview/251795411/E8E548F5E9794CFBPQ/1?accountid=13965

Smith, R. M. (1994). A comparison of the power of Rasch total and between-item fit statistics to detect measurement disturbances. *Educational and Psychological Measurement, 54*(1), 42-55.

Smith, R. M. (1996). A comparison of the Rasch separate calibration and between-fit methods of detecting item bias. *Educational and Psychological Measurement, 56*(3), 403-418.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1*(2), 199-218.

Smith, R. M. (2004a). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 73-92). Maple Grove, MN: JAM Press.

Smith, R. M. (2004b). Detecting item bias with the Rasch model. *Journal of Applied Measurement, 5*(4), 430-449.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

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

South Carolina Arts Assessment Program (2003-2020) [Measurement instrument]. Columbia, SC: The SC Department of Education (SCDE).

South Carolina College- and Career-Ready Assessments (2016-2020) [Measurement instrument]. Columbia, SC: The SC Department of Education (SCDE).

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*(2), 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology*, *5*(4), 417-426.

Stagnaro-Green, A. S., & Downing, S. M. (2006). Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. *Medical Teacher*, *28*(6), 566-568.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758-765. Retrieved on October 29, 2019 from https://journals.sagepub.com/doi/pdf/10.1177/003172170208301010

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, *67*(4), 485-518.

Su, Y.-H., Sheu, C.-F., & Wang, W.-C. (2007). Computing confidence intervals of item fit statistics in the family of Rasch models using the bootstrap method. *Journal of Applied Measurement*, *8*(2), 190-203.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Sykes, R. C., Yen, W., & Ito, K. (1996, April). *Scaling polytomous items that have been scored by two raters*. Paper presented at the annual meeting of NCME, New York, NY.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*(2), 198-206.

Test of English as a Foreign Language (1964-2020) [Measurement instrument]. Princeton, NJ: English Testing Service (ETS).

Tutz, G., & Berger, M. (2016). Item-focused trees for the identification of items in differential item functioning. *Psychometrika, 81*(3), 727-750.

U.S. Congress Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, D.C.: U. S. Government Printing Office. Retrieved on May 9, 2019 from http://govinfo.library.unt.edu/ota/Ota_1/DATA/1992/9236.PDF

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*(2), 123-140.

Viger, S. G., Wolfe, E. W., Dozier, H., & Machtmes, K. (2006). Validation of a questionnaire used to assess safety and professionalism among arborists. *Journal of Applied Measurement, 7*(3), 292-306.

Vista, A. (2016). The role of PS ability and RC skill in predicting growth trajectories of mathematics achievement. *Educational Assessment and Evaluation, 3*(1), 1-22. Retrieved on January 18, 2019 from https://www.cogentoa.com/article/10.1080/2331186X.2016.1222720.pdf

von Davier, M., & Xu, X. (2011). Measuring growth in a longitudinal large-scale assessment within general latent variable model. *Psychometrika, 76*(2), 318-336.

Walker, A. A., Jennings, J. K., & Engelhard, G. (2018). Using person response functions to investigate areas of person misfit related to item characteristics. *Educational Assessment, 23*(1), 47-68.

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment, 29*(4), 364-376. doi: 10.1177/0734282911406666

Wang, W.-C., & Chen, C.-T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404.

Wind, S. A., & Engelhard, G. (2012). Examining rating quality in writing assessment: Rater agreement, error, and accuracy. *Journal of Applied Measurement, 13*(4), 321-335.

Wolfe, E. W. (2013). A bootstrap approach to evaluating person and item fit to the Rasch model. *Journal of Applied Measurement, 14*(1), 1-9.

Wolfe, E. W., & Chiu, C. W. T. (1997, March). *Detecting rater effects with a multi-faceted rating scale model*. Paper presented at the annual meeting of NCME, Chicago, IL.

Wolfe, E. W., & Miller, T. R. (1997). Barriers to the implementation of portfolio assessment in secondary education. *Applied Measurement in Education, 10*(3), 235-251.

Wong, E., Woo, J. W., Hui, E., & Ho, S. C. (2004). Examination of the Philadelphia Geriatric Morale Scale as a subjective quality-of-life measure in elderly Hong Kong Chinese. *The Gerontologist, 44*(3), 408-417.

Wright, B. D. (1980). Afterword. In G. Rasch, *Probabilistic models for some intelligence and attainment tests* 1960/1980. Chicago, IL: University of Chicago Press.

Wright, B. D. (1989). Dichotomous Rasch Model derived from counting right answers: Raw scores as sufficient statistics. *Rasch Measurement Transactions*, *3*(2), 62. Retrieved on July 5, 2019 from https://www.rasch.org/rmt/rmt32e.htm

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement*, *1*, 281-295.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370. Retrieved on February 21, 2019 from http://www.rasch.org/rmt/rmt83b.htm

Wright, B. D., Linacre, J. M., & Schultz, M. (1989). *A user's guide to BIGSCALE: Rasch model rating scale analysis computer program* (Version 1.5). Chicago, IL: MESA Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Wright, B. D., & Masters, G. N. (1990). Computation of OUTFIT and INFIT statistics. *Rasch Measurement Transactions*, *3*(4), 84-85. Retrieved on December 5, 2018 from https://www.rasch.org/rmt/rmt34e.htm

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23-48.

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

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15-27.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*(4), 339-355.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers: Theory into practice*. Singapore: Springer Nature.

Wyse, A. E. (2011). The similarity of bookmark cut scores with different response probability values. *Educational and Psychological Measurement, 71*(6), 963-985.

Yan, J. W. (1997, March). *Examining local item dependence effects in a large-scale science assessment by a Rasch partial credit model*. Paper presented at the annual meeting of AERA, Chicago, IL.

Yen, W. M., & Henderson, D. L. (2002). Professional standards related to using large-scale state assessments in decisions for individual students. *Measurement and Evaluation in Counselling and Development, 35*, 132-143.

Yildirim, H. H., Yildirim, S., & Verhelst, N. (2014). Profile analysis as a generalized differential item functioning analysis method. *Education and Science, 39*(172), 49-64.

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Hillsdale, NJ: Lawrence Erlbaum.

Zieky, M. (2003). *A DIF Primer*. Princeton, NJ: Educational Testing Service. Retrieved on July 22, 2019 from https://www.ets.org/s/praxis/pdf/dif_primer.pdf

Zumbo, B. D. (1999). *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for*

binary and Likert-type (ordinal) item scores (Report). Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* (RR-12-08). Princeton, NJ: Education Testing Service. Retrieved on January 7, 2020 from https://www.ets.org/Media/Research/pdf/RR-12-08.pdf

APPENDIX A

SELECTED Winsteps® CONTROL FILES AND LOOP FILE

```
&INST
TITLE = "Dissertation"
PERSON = Person ; persons are ...
ITEM = Item ; items are ...
ITEM1 = 1 ; column of response to first item in data record
NI = 50 ; number of items
NAME1 = 52 ; column of first character of person identifying label
NAMELEN = 4 ; length of person label
XWIDE = 1 ; number of columns per item response
CODES = 01 ; valid codes in data file
UIMEAN = 0 ; item mean for local origin
USCALE = 1 ; user scaling for logits
UDECIM = 2 ; reported decimal places for user scaling
TFILE=*
14.1        ; output table 14.1
*
BOXSHOW = NO; output no border
FITI = 0; include all items
&END
END LABELS
```

```
C:

MKDIR \Users\Yin\Desktop\I50P5000

CD \I50P5000

REM

set /a test=1

:loop

REM

START /WAIT C:\Winsteps\winsteps.exe BATCH=YES Control.txt
DATA%test%.out.txt data=Data%test%.txt

REM

START /WAIT C:\Winsteps\winsteps.exe BATCH=YES Control.txt
data=Data%test%.txt DATA%test%.out.txt TFILE=* 14.1 *

REM

set /a test=%test%+1

if not "%test%"=="1001" goto loop

PAUSE
```

```
&INST
ITEM1 = 1 ; Starting column of item responses
NI = 50 ; Number of items
NAME1 = 51 ; Starting column for person label in data record
NAMLEN = 3 ; Length of person label
XWIDE = 1 ; Matches the widest data value observed
; GROUPS = 0 ; Partial Credit model: in case items have different rating scales
CODES = 01 ; matches the data
TOTALSCORE = Yes ; Include extreme responses in reported scores
; Person Label variables: columns in label: columns in line
@Group = $C51W3 ; $C51W3
DIF=@Group  ;Use the Group for DIF detection
PSUBTOTAL = @Group      ; Subtotal by group
TFILE=*
30.4     ; output table 30.4
*
BOXSHOW = NO; output no border
FITI = 0; include all items

&END

END LABELS
```

APPENDIX B

SELECTED SAS® CODES TO SUMMARIZE FIT INDICES

```
filename indata pipe 'dir C:\Users\burgessy\Desktop\SIMResults\50Items\I50P5000 /b';


data file_list; length fname $20;
infile indata truncover;          /* infile statement for file names */
input fname $20.;       /* read the file names from the directory */
call symput ('num_files',_n_); /* store the record number in a macro variable */
run;
%macro fileread;

%do j=1 %to &num_files;
data _null_;
set file_list;
if _n_=&j;
call symput ('filein',fname);
run;


/* read the data lines into a temporary file */

data temp;
infile "C:\Users\burgessy\Desktop\SIMResults\50Items\I50P5000\&filein" firstobs=12
obs=61 missover;
input var1-var13 var14 $;
run;

/* assemble the individual files */

%if &j=1 %then %do;
data data_all;
set temp;
run;
```

```
%end;

%else %do;
data data_all;
set data_all
temp;
run;
%end;

%end;  /* end of do-loop with index j */

%mend fileread;
%fileread;

data data_all (rename=(var1=entrynumber var2=score var3=totalcount var4=measure
var5= modelse var6=infitmnsq
var7=infitZSTD var8=outfitmnsq var9=outfitZSTD var10=CORR var11=CORRExp
var12=ExtObsPerc var13=MatchExpPerc var14=Item));
set data_all;
run;

proc sort data=data_all;
by item;
run;

data data_allrep;
set data_all;
by entrynumber;
if first.entrynumber
then replicate=1;
else replicate+1;
run;

proc sort data=data_allrep;
by replicate item;
run;

proc means data=data_allrep;
class replicate;
var infitmnsq infitZSTD outfitmnsq outfitZSTD;
output out=data_allrep2 mean=replicate_mean;
run;

data data_allreppct;
set data_allrep;
by entrynumber replicate;
```

```
if infitmnsq>1.3
then infitmnsq3=1;
else infitmnsq3=0;
if infitmnsq>1.2
then infitmnsq2=1;
else infitmnsq2=0;
if infitmnsq>1.1
then infitmnsq1=1;
else infitmnsq1=0;

if infitmnsq<0.9
then infitmnsq9=1;
else infitmnsq9=0;
if infitmnsq<0.8
then infitmnsq8=1;
else infitmnsq8=0;
if infitmnsq<0.7
then infitmnsq7=1;
else infitmnsq7=0;

if outfitmnsq>1.3
then outfitmnsq3=1;
else outfitmnsq3=0;
if outfitmnsq>1.2
then outfitmnsq2=1;
else outfitmnsq2=0;
if outfitmnsq>1.1
then outfitmnsq1=1;
else outfitmnsq1=0;

if outfitmnsq<0.9
then outfitmnsq9=1;
else outfitmnsq9=0;
if outfitmnsq<0.8
then outfitmnsq8=1;
else outfitmnsq8=0;
if outfitmnsq<0.7
then outfitmnsq7=1;
else outfitmnsq7=0;

if infitZSTD>4.0
then infitZSTD4=1;
else infitZSTD4=0;
if infitZSTD>3.0
then infitZSTD3=1;
```

```
else infitZSTD3=0;
if infitZSTD>2.0
then infitZSTD2=1;
else infitZSTD2=0;

if infitZSTD<-4.0
then infitZSTD4neg=1;
else infitZSTD4neg=0;
if infitZSTD<-3.0
then infitZSTD3neg=1;
else infitZSTD3neg=0;
if infitZSTD<-2.0
then infitZSTD2neg=1;
else infitZSTD2neg=0;

if outfitZSTD>4.0
then outfitZSTD4=1;
else outfitZSTD4=0;
if outfitZSTD>3.0
then outfitZSTD3=1;
else outfitZSTD3=0;
if outfitZSTD>2.0
then outfitZSTD2=1;
else outfitZSTD2=0;

if outfitZSTD<-4.0
then outfitZSTD4neg=1;
else outfitZSTD4neg=0;
if outfitZSTD<-3.0
then outfitZSTD3neg=1;
else outfitZSTD3neg=0;
if outfitZSTD<-2.0
then outfitZSTD2neg=1;
else outfitZSTD2neg=0;

run;

proc sort data=data_allreppct;
by replicate item;
run;
proc sql;
create table infitmnsq as
select   mean(infitmnsq3) as infitmnsq3_pct,
         mean(infitmnsq2) as infitmnsq2_pct,
         mean(infitmnsq1) as infitmnsq1_pct,
         mean(infitmnsq7) as infitmnsq7_pct,
```

```
        mean(infitmnsq8) as infitmnsq8_pct,
        mean(infitmnsq9) as infitmnsq9_pct
from work.data_allreppct
group by replicate
order by replicate;
quit;

proc means;
var infitmnsq3_pct infitmnsq2_pct infitmnsq1_pct infitmnsq7_pct infitmnsq8_pct
infitmnsq9_pct;
run;

proc sql;
create table infitZSTD as
select  mean(infitZSTD4) as infitZSTD4_pct,
        mean(infitZSTD3) as infitZSTD3_pct,
        mean(infitZSTD2) as infitZSTD2_pct,
        mean(infitZSTD2neg) as infitZSTD2neg_pct,
        mean(infitZSTD3neg) as infitZSTD3neg_pct,
        mean(infitZSTD4neg) as infitZSTD4neg_pct
from work.data_allreppct
group by replicate
order by replicate;
quit;

proc means;
var infitZSTD4_pct infitZSTD3_pct infitZSTD2_pct infitZSTD2neg_pct
infitZSTD3neg_pct infitZSTD4neg_pct;
run;

proc sql;
create table outfitmnsq as
select  mean(outfitmnsq3) as outfitmnsq3_pct,
        mean(outfitmnsq2) as outfitmnsq2_pct,
        mean(outfitmnsq1) as outfitmnsq1_pct,
        mean(outfitmnsq7) as outfitmnsq7_pct,
        mean(outfitmnsq8) as outfitmnsq8_pct,
        mean(outfitmnsq9) as outfitmnsq9_pct
from work.data_allreppct
group by replicate
order by replicate;
quit;

proc means;
var outfitmnsq3_pct outfitmnsq2_pct outfitmnsq1_pct outfitmnsq7_pct outfitmnsq8_pct
outfitmnsq9_pct;
```

```
run;

proc sql;
create table outfitZSTD as
select mean(outfitZSTD4) as outfitZSTD4_pct,
        mean(outfitZSTD3) as outfitZSTD3_pct,
        mean(outfitZSTD2) as outfitZSTD2_pct,
        mean(outfitZSTD2neg) as outfitZSTD2neg_pct,
        mean(outfitZSTD3neg) as outfitZSTD3neg_pct,
        mean(outfitZSTD4neg) as outfitZSTD4neg_pct
from work.data_allreppct
group by replicate
order by replicate;
quit;

proc means;
var outfitZSTD4_pct outfitZSTD3_pct outfitZSTD2_pct outfitZSTD2neg_pct
outfitZSTD3neg_pct outfitZSTD4neg_pct;
run;

data data_allcv;
set data_allreppct;
lcv1=infitMNSQ-2*modelse;
ucv1=infitMNSQ+2*modelse;
lcv2=outfitMNSQ-2*modelse;
ucv2=outfitMNSQ+2*modelse;
run;

proc sql;
create table criticalvalues as
select lcv1 as lcvinfit,
        ucv1 as ucvinfit,
        lcv2 as lcvoutfit,
        ucv2 as ucvoutfit
from work.data_allcv
group by replicate
order by replicate;
quit;

proc means;
var lcvinfit ucvinfit lcvoutfit ucvoutfit;
run;
```

APPENDIX C

SELECTED R CODES TO GENERATE RESPONSES FOR POWER ANALYSIS

```
# Read in the item parameters #

source("http://people.stat.sc.edu/habing/courses/irtS14.txt")


############ Total number of items ################

nitem <- 50

###### Number of DIF items ##########

ndifitem_1 <- 0.04*50

ndifitem_2 <- 0.1*50

ndifitem_3 <- 0.2*50

ndifitem_4 <- 0.4*50

######### Sample size ############

N_1 <- 5000

N_2 <- 10000

N_3 <- 25000

N_4 <- 50000
```

```
############ Create ability for two groups (A,B) and item difficulty parameter
##########

set.seed(2019)

ability_1 <- rnorm(N_1,mean=0,sd=1)

ability_1A <- ability_1[1:2500]

ability_1B <- ability_1[2501:5000]

ability_2 <- rnorm(N_2,mean=0,sd=1)

ability_2A <- ability_2[1:5000]

ability_2B <- ability_2[5001:10000]

ability_3 <- rnorm(N_3,mean=0,sd=1)

ability_3A <- ability_3[1:12500]

ability_3B <- ability_3[12501:25000]

ability_4 <- rnorm(N_4,mean=0,sd=1)

ability_4A <- ability_4[1:25000]

ability_4B <- ability_4[25001:50000]

difficulty<-runif(nitem,-2,2)

# difficulty<-uniform(30,min=-2,max=2)


# Create DIF items. Here we are going to create 2,5,10,20 large DIF items ######

dif_2l <- difficulty

dif_2l[49] <- dif_2l[49]+0.65

dif_2l[50] <- dif_2l[50]+0.65


dif_5l <- difficulty
```

```
dif_5l[46] <- dif_5l[46]+0.65
dif_5l[47] <- dif_5l[47]+0.65
dif_5l[48] <- dif_5l[48]+0.65
dif_5l[49] <- dif_5l[49]+0.65
dif_5l[50] <- dif_5l[50]+0.65

dif_10l <- difficulty
dif_10l[41] <- dif_10l[41]+0.65
dif_10l[42] <- dif_10l[42]+0.65
dif_10l[43] <- dif_10l[43]+0.65
dif_10l[44] <- dif_10l[44]+0.65
dif_10l[45] <- dif_10l[45]+0.65
dif_10l[46] <- dif_10l[46]+0.65
dif_10l[47] <- dif_10l[47]+0.65
dif_10l[48] <- dif_10l[48]+0.65
dif_10l[49] <- dif_10l[49]+0.65
dif_10l[50] <- dif_10l[50]+0.65

dif_20l <- difficulty
dif_20l[31] <- dif_20l[31]+0.65
dif_20l[32] <- dif_20l[32]+0.65
dif_20l[33] <- dif_20l[33]+0.65
dif_20l[34] <- dif_20l[34]+0.65
dif_20l[35] <- dif_20l[35]+0.65
dif_20l[36] <- dif_20l[36]+0.65
```

```
dif_20l[37] <- dif_20l[37]+0.65
dif_20l[38] <- dif_20l[38]+0.65
dif_20l[39] <- dif_20l[39]+0.65
dif_20l[40] <- dif_20l[40]+0.65
dif_20l[41] <- dif_20l[41]+0.65
dif_20l[42] <- dif_20l[42]+0.65
dif_20l[43] <- dif_20l[43]+0.65
dif_20l[44] <- dif_20l[44]+0.65
dif_20l[45] <- dif_20l[45]+0.65
dif_20l[46] <- dif_20l[46]+0.65
dif_20l[47] <- dif_20l[47]+0.65
dif_20l[48] <- dif_20l[48]+0.65
dif_20l[49] <- dif_20l[49]+0.65
dif_20l[50] <- dif_20l[50]+0.65

#Generate One Simulated Data #
source("http://people.stat.sc.edu/habing/courses/irtS14.txt")
group1<-c(rep("A",2500),rep("B",2500))
group2<-c(rep("A",5000),rep("B",5000))
group3<-c(rep("A",12500),rep("B",12500))
group4<-c(rep("A",25000),rep("B",25000))

nsims<-1000

simdata <- function(ability1,ability2,DIFitem,group){
```

```
 for (sim in 1:nsims){
   U1<- irtgen(ability1,A=1,B=difficulty,rep(0,50),type="norm")
   U2<-irtgen(ability2,A=1,DIFitem,rep(0,50),type="norm")
   U<- rbind(U1,U2)
   U<- data.frame(U,group)
   assign(paste0("SimData",sim),U)
   filename1 <- paste("SimData",sim,".txt",sep="")
   write.table(get(paste0("SimData",sim)),file=filename1,sep="",row.name=F,col.names
= F)
  }
}

# 2500 sample size each group, large DIF,1DIF items####
simdata(ability_1A,ability_1B,dif_2l,group1)
# 2500 sample size each group, large DIF,3DIF items ####
simdata(ability_1A,ability_1B,dif_5l,group1)
# 2500 sample size each group, large DIF,6DIF items ####
simdata(ability_1A,ability_1B,dif_10l,group1)
# 2500 sample size each group, large DIF,12DIF items ####
simdata(ability_1A,ability_1B,dif_20l,group1)

# 5000 sample size each group, large DIF,1DIF items####
simdata(ability_2A,ability_2B,dif_2l,group2)
# 5000 sample size each group, large DIF,3DIF items ####
simdata(ability_2A,ability_2B,dif_5l,group2)
```

```
# 5000 sample size each group, large DIF,6DIF items ####
simdata(ability_2A,ability_2B,dif_10l,group2)
# 5000 sample size each group, large DIF,12DIF items ####
simdata(ability_2A,ability_2B,dif_20l,group2)

# 12500 sample size each group, large DIF,1DIF items####
simdata(ability_3A,ability_3B,dif_2l,group3)
# 12500 sample size each group, large DIF,3DIF items ####
simdata(ability_3A,ability_3B,dif_5l,group3)
# 12500 sample size each group, large DIF,6DIF items ####
simdata(ability_3A,ability_3B,dif_10l,group3)
# 12500 sample size each group, large DIF,12DIF items ####
simdata(ability_3A,ability_3B,dif_20l,group3)

# 25000 sample size each group, large DIF,1DIF items####
simdata(ability_4A,ability_4B,dif_2l,group4)
# 25000 sample size each group, large DIF,3DIF items ####
simdata(ability_4A,ability_4B,dif_5l,group4)
# 25000 sample size each group, large DIF,6DIF items ####
simdata(ability_4A,ability_4B,dif_10l,group4)
# 25000 sample size each group, large DIF,12DIF items ####
simdata(ability_4A,ability_4B,dif_20l,group4)
```

APPENDIX D

Winsteps® SAMPLE OUTPUT FOR BETWEEN-ITEM FIT STATISTICS

TABLE 30.4

INPUT: 50000 PERSON 25 ITEM REPORTED: 50000 PERSON 25 ITEM 2 CATS WINSTEPS 4.4.8

DIF class/group specification is: DIF=@Group

PERSON CLASSES	SUMMARY DIF CHI-SQUARED	D.F.	PROB.	BETWEEN-CLASS/GROUP ITEM UNWTD MNSQ	ZSTD	Number	Name
2	63.7042	1	.0000	63.9292	6.83	1	I0001
2	29.1739	1	.0000	28.9890	4.87	2	I0002
2	20.6326	1	.0000	20.4847	4.15	3	I0003
2	32.9327	1	.0000	33.2318	5.17	4	I0004
2	71.2943	1	.0000	71.6271	7.16	5	I0005
2	5.5372	1	.0186	5.6665	2.13	6	I0006
2	88.1471	1	.0000	88.8114	7.81	7	I0007
2	71.3264	1	.0000	72.0159	7.18	8	I0008
2	55.0539	1	.0000	55.2620	6.43	9	I0009
2	32.4286	1	.0000	32.8350	5.14	10	I0010
2	64.9189	1	.0000	64.4907	6.86	11	I0011
2	23.3563	1	.0000	23.6806	4.44	12	I0012
2	54.9787	1	.0000	54.2604	6.38	13	I0013
2	64.6358	1	.0000	64.4295	6.85	14	I0014
2	36.9369	1	.0000	36.6548	5.40	15	I0015
2	42.7445	1	.0000	42.3203	5.74	16	I0016
2	42.1573	1	.0000	42.4552	5.75	17	I0017
2	79.6139	1	.0000	80.2599	7.50	18	I0018
2	39.3513	1	.0000	39.2892	5.56	19	I0019
2	86.9718	1	.0000	86.2351	7.72	20	I0020
2	53.5230	1	.0000	54.3562	6.39	21	I0021
2	53.0248	1	.0000	52.4575	6.29	22	I0022
2	65.3380	1	.0000	64.6835	6.87	23	I0023
2	13.7623	1	.0002	13.8027	3.44	24	I0024
2	1850.5184	1	.0000	1903.2266	24.64	25	I0025